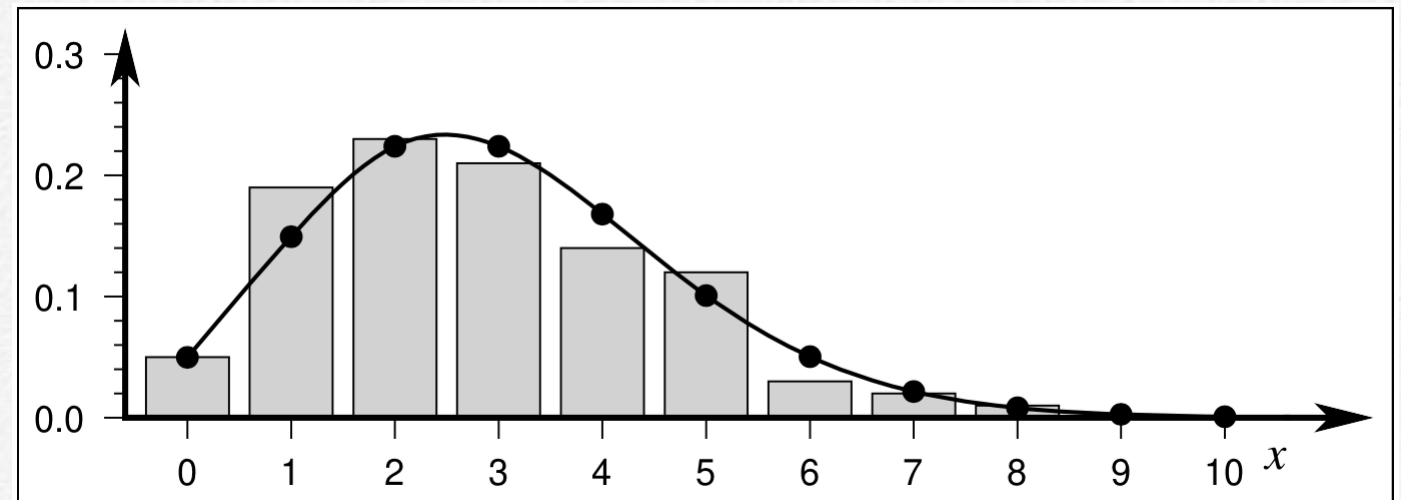


STATISTICAL CONCEPTS

We make a histogram of the data by normalizing the observed frequencies by the total count and superimposing the Poisson distribution for the expected rate. The result (Figure 3.12) shows a very good fit.

Figure 3.12: Histogram of observed decay rate frequencies (bars) and the theoretical Poisson distribution (circles) for the expected rate $\lambda = 3$.



3.4 Continuous Probability Distributions

While many populations are of a discrete nature (e.g., outcomes of coin tosses, numbers of microfossils in a core, etc.), we are very often dealing with observations of a phenomenon that can take on any of a continuous spectrum of values. We may sample the phenomenon at certain points in space-time and thus have discrete observations. Nevertheless, the underlying probability distribution is continuous (e.g., Figure 3.13).

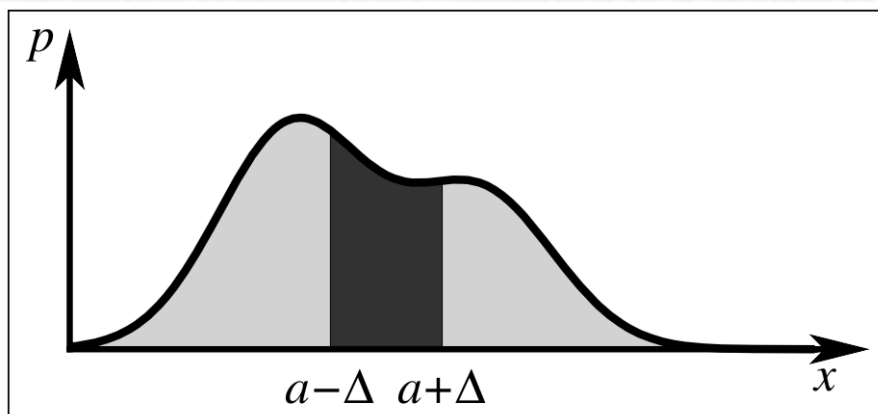


Figure 3.13: Example of a continuous probability density function (pdf). The area under any pdf must equal 1. The finite probability identified in (3.75) is indicated in dark gray.

Continuous distributions can be thought of as the limit for discrete distributions when the “spacing” between events shrinks to zero.

STATISTICAL CONCEPTS

Hence, we must replace the summation in (3.68) with the integral

$$\int_{-\infty}^{\infty} p(x) dx = 1. \quad (3.74)$$

Because of their continuous nature, functions such as $p(x)$ in (3.74) are called *probability density functions* (pdf). The probability of an event is still defined by the area under the curve, but now we must integrate to find the area and hence the probability. E.g., the probability that a random variable will take on a value between $a - \Delta$ and $a + \Delta$ is

$$P(a \pm \Delta) = \int_{a-\Delta}^{a+\Delta} p(x) dx. \quad (3.75)$$

As $\Delta \rightarrow 0$ we find that the probability goes to zero. Thus, the probability of getting exactly $x = a$ is nil.

The *cumulative distribution function* (cdf) gives the probability that an observation less than or equal to a will occur. We obtain the integral expression for this distribution by replacing the lower limit by $-\infty$ and the upper limit by a , finding

$$P_c(a) = \int_{-\infty}^a p(x) dx. \quad (3.76)$$

Obviously, as $a \rightarrow \infty$, $P_c(a) \rightarrow 1$. Given the cumulative distribution function we can revisit (3.75) and instead state

$$P(a \pm \Delta) = P_c(a + \Delta) - P_c(a - \Delta). \quad (3.77)$$

3.4.1 The normal distribution

So far the function $p(x)$ has been arbitrary. Any continuous function with unit area under the curve (i.e., 3.74) would qualify. We will now turn our attention to the best known and most frequently used pdf: the *normal distribution*. Its study dates back to 18th century investigations into the nature of experimental error. It was found that repeat measurements of the same quantity displayed a surprising degree of regularity.

STATISTICAL CONCEPTS

In particular, the German scientist K. F. Gauss played a major role in developing the theoretical foundations for the normal distribution, hence its other name: the **Gaussian distribution**. It is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad (3.78)$$

where μ and σ have been defined previously. The constant term before the exponential normalizes the area under the curve to unity (Figure 3.14). As discussed in Section 3.2.4, it is often convenient to transform your data into so-called **standard scores**:

$$z_i = \frac{x_i - \mu}{\sigma}, \quad (3.79)$$

in which case (3.78) reduces to

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad (3.80)$$

which has zero mean and unit standard deviation.

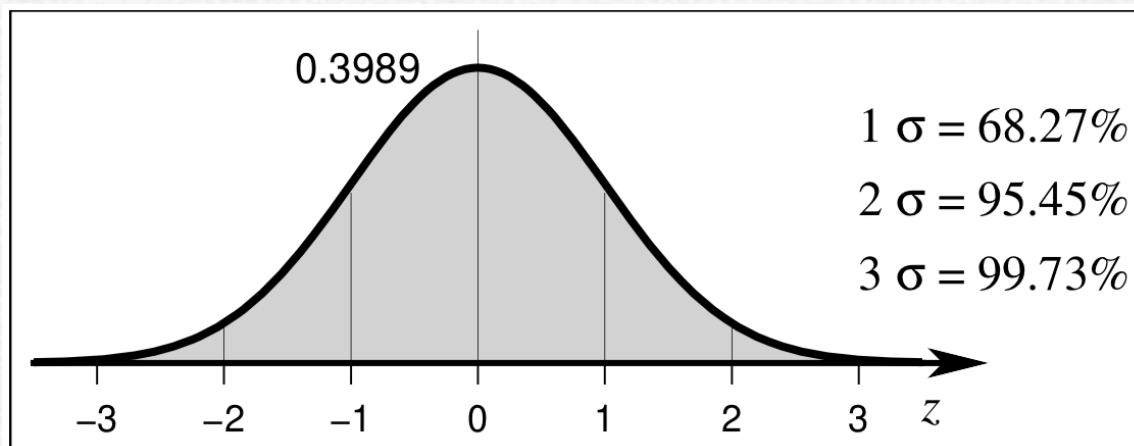


Figure 3.14: A normally distributed data set will have almost all of its values within $\pm 3\sigma$ of the mean (this corresponds to 99.73% of the data; see legend for percentages corresponding to other multiples of $\pm\sigma$).

Given the functional form of $p(z)$ we can evaluate the probability that an observation z will be $\leq a$:

$$P_c(a) = \int_{-\infty}^a p(z) dz = \int_{-\infty}^0 p(z) dz + \int_0^a p(z) dz = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^a e^{-\frac{z^2}{2}} dz. \quad (3.81)$$

STATISTICAL CONCEPTS

Let

$$u^2 = \frac{z^2}{2}, \text{ hence } dz = \sqrt{2}du, \quad \operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad (3.82)$$

then

$$P_c(a) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\frac{a}{\sqrt{2}}} e^{-u^2} du = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \frac{\sqrt{\pi}}{2} \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) \right]. \quad (3.83)$$

It follows that, for any value z , the cumulative distribution function is

$$P_c(z) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) \right]. \quad (3.84)$$

Here, erf represents the *error function* and it is defined by the definite integral in (3.83) and tabulated in Table A.1. Furthermore, the probability that z falls between a and b must necessarily be

$$P_c(a \leq z \leq b) = P_c(b) - P_c(a) = \frac{1}{2} \left[\operatorname{erf}\left(\frac{b}{\sqrt{2}}\right) - \operatorname{erf}\left(\frac{a}{\sqrt{2}}\right) \right]. \quad (3.85)$$

Example 3–11. Investigations into the strength of olivine have provided estimates of Young's modulus (E) that follow a normal distribution given by $\mu = 1.0 \cdot 10^{11}$ Pa and $\sigma = 1.0 \cdot 10^{10}$ Pa. What is the probability that a single estimate E will lie in the interval $9.8 \cdot 10^{10}$ Pa $< E < 1.1 \cdot 10^{11}$ Pa? We convert the limits to normal scores and find they correspond to the interval $-0.2 \leq z \leq 1.0$. Using these values for a and b in (3.85) (or using Table A.1) we find the probability to be 0.4206.

Approximate binomial distribution

Like the Poisson distribution, the normal distribution may also serve as an approximation to the binomial distribution when n is large. More specifically, this approximation holds when both np and $(1 - p)n$ exceed 5. Under those circumstances, the mean and standard deviation of the approximate normal distribution become

STATISTICAL CONCEPTS

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}, \quad (3.86)$$

leading to the simplified distribution

$$P_b(x) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp \left\{ \frac{-(x-np)^2}{2np(1-p)} \right\}. \quad (3.87)$$

Example 3–12. What is the probability that at least 70 of 100 sand grains will be larger than 0.5 mm if the probability that any single grain is that large is $p = 0.75$? Using the approximation (3.86) we find $\mu = np = 75$ and $s = \sqrt{np(1-p)} = 4.33$. Converting 69.5 (halfway between 69 and 70) to a z score gives -1.27 , and we find via Table A.1 that the probability becomes 0.898 or about 90%.

3.4.2 The exponential distribution

Another important probability density distribution is the *exponential* distribution. It is given by

$$p_e(x) = \lambda e^{-\lambda x} \quad (3.88)$$

for some constant λ . However, most of the time we will see it used as a cumulative distribution function:

$$P_c(x) = 1 - e^{-\lambda x}. \quad (3.89)$$

Eq. (3.89) gives the probability that the observation a will be in the range $0 \leq a \leq x$.

Example 3–13. It has been reported that the heights (z) of Pacific seamounts follow an exponential distribution defined as

$$P_c(z \leq h) = 1 - e^{-h/340}, \quad (3.90)$$

which gives the probability that a seamount is shorter than h meters. Equation (3.90) then predicts that we might expect that

$$P_c(1000) = 1 - e^{-1000/340} \approx 95\% \quad (3.91)$$

of them are less than one km tall.

STATISTICAL CONCEPTS

3.4.3 Log-normal distribution

Many data sets, such as grain-sizes of sediments and geochemical concentrations, have very skewed and long-tailed distributions (e.g., Figure 3.15). In general, such distributions arise when the observed quantities have errors that depend on *products* rather than *sums*. It therefore follows that the *logarithm* of the data may be normally distributed. Hence, taking the logarithm of your data may make the transformed distribution look normal. If this is the case, you can apply standard statistical techniques applicable to normal distributions to the logarithm of your data and convert the results (e.g., mean, standard deviation) back to get the proper units. The log-normal probability density distribution is therefore given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log x - \mu}{\sigma}\right)^2}. \quad (3.92)$$

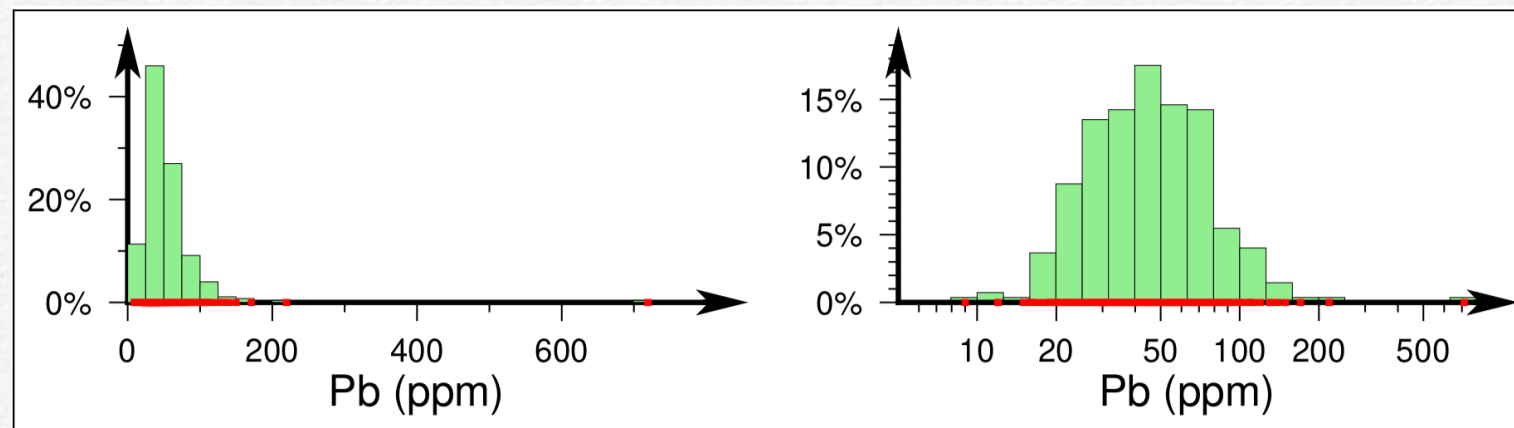


Figure 3.15: (left) The concentration of Pb in soil is very long-tailed and clearly not normally distributed. The red squares indicate individual sample values. (right) The same distribution after taking the logarithm of the data values. The resulting distribution is approximately normal, hence a log-normal distribution might be suitable to describe the data.

3.5 Inferences about Means

The central limits theorem states that the mean of a large sample taken from any distribution will be normally distributed even if the data themselves are not normally distributed, and furthermore it says that the sample mean is an unbiased estimator of the population mean. We can then use our knowledge of the normal distribution to quantify our faith in the precision of our sample mean. We already know that $s_{\bar{x}} = \sigma/\sqrt{n}$, so we can state with probability $1 - \alpha$ that \bar{x} will differ from μ by at most E , which is given by

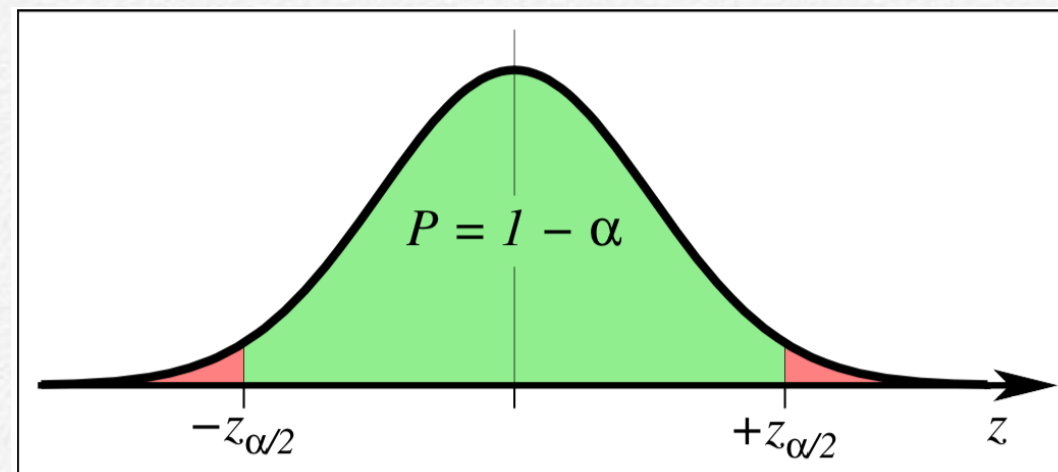
STATISTICAL CONCEPTS

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \quad (3.93)$$

where s is our estimate of σ . In other words, the chance that \bar{x} exceeds the $\pm z_{\alpha/2}$ confidence interval is α . These error estimates apply to large samples ($n \geq 30$) and infinite populations. In those cases we can use our sample standard deviation s in place of σ , which we usually do not know. Here, (3.93) can be inverted to yield the sample size necessary to be confident that the error in our sample mean is no larger than E , and we find

$$n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2. \quad (3.94)$$

Figure 3.16: Probability is α that a value will fall in one of the two tails of the normal distribution, and $\alpha/2$ that it will fall in a specific tail.



The normal score for our sample mean is

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}. \quad (3.95)$$

Since this statistic is normally distributed we know that the probability is $1 - \alpha$ that z will take on a value in the interval $-z_{\alpha/2} < z < +z_{\alpha/2}$. Plugging in for the limits on z ,

$$-z_{\alpha/2} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < +z_{\alpha/2} \quad (3.96)$$

STATISTICAL CONCEPTS

or

$$\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}. \quad (3.97)$$

Rearranging, we find

$$\mu = \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}. \quad (3.98)$$

Eq. (3.98) shows the *confidence interval* on μ at the $1 - \alpha$ confidence level. Very often, our confidence levels will be 95% ($\sim 2\sigma$) or 99% ($\sim 3\sigma$).

3.5.1 Small samples

The previous section dealt with large ($n \geq 30$) samples, where we could assume that \bar{x} would be normally distributed as dictated by the central limits theorem. For smaller samples we must assume instead that the *population we are sampling* is normally distributed. We can then base our inferences on the statistic

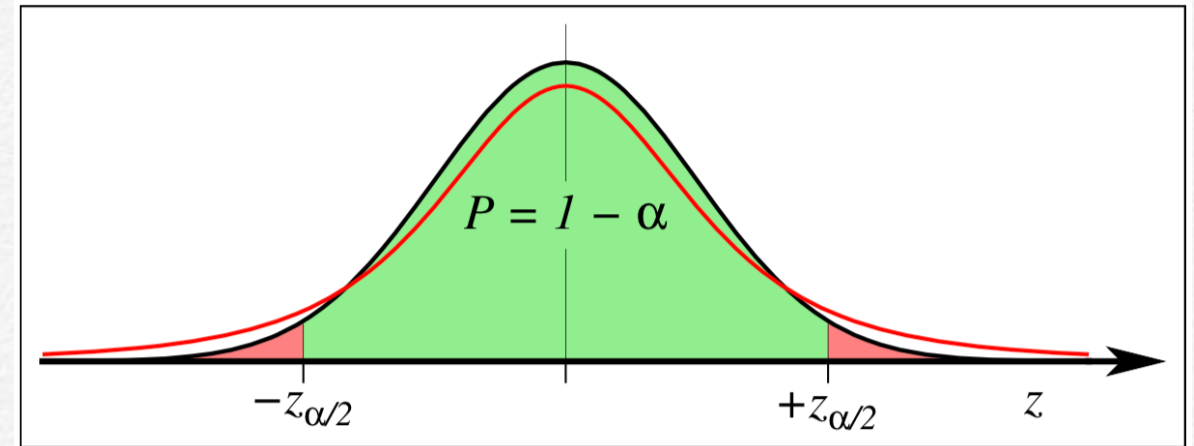
$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad (3.99)$$

whose distribution is called the *Student's t*-distribution (Figure 3.17). It is similar to the normal distribution but its shape depends on the degrees of freedom, $\nu = n - 1$. For large n (and hence ν) the t statistics approach the z statistics. As for z statistics, one can find tables with t values for various combinations of confidence levels and degrees of freedom (see Table A.2). For insight into what the t -distribution and others really are, get *Numerical Recipes* by Press et al. This excellent book gives both theory and computer code (in C++, C, FORTRAN, Java and a host of legacy languages).

STATISTICAL CONCEPTS

Example 3–14. Given our sandstone density estimates from earlier, i.e., {2.2, 2.25, 2.25, 2.3, 2.3, 2.3, 2.35}, what is the 95% confidence interval on the population mean?

Figure 3.17: The same normal distribution and critical tails as in Figure 3.16, overlain by the Student's t -distribution for $v = 3$ degrees of freedom (red line). For small samples the probability distribution becomes wider.



Answer: We have $\bar{x} = 2.28$ with $s = 0.05$, and $\alpha = 1 - 95\% = 0.05$. The degrees of freedom $v = n - 1 = 6$. Table A.2 gives $t_{\alpha/2, v} = t_{0.025, 6} = 2.447$. Using (3.97), we find our sample mean brackets the population mean, thus (with $t_{\alpha/2}$ instead of $z_{\alpha/2}$ and s instead of σ)

$$2.28 - 2.447 \cdot \frac{0.05}{\sqrt{7}} < \mu < 2.28 + 2.447 \cdot \frac{0.05}{\sqrt{7}} \quad (3.100)$$

or (since the bounds are symmetrical)

$$2.234 < \mu < 2.326 \quad (3.101)$$

or

$$\mu = 2.280 \pm 0.046. \quad (3.102)$$

STATISTICAL CONCEPTS

A.1 Cumulative Probabilities for the Normal Distribution

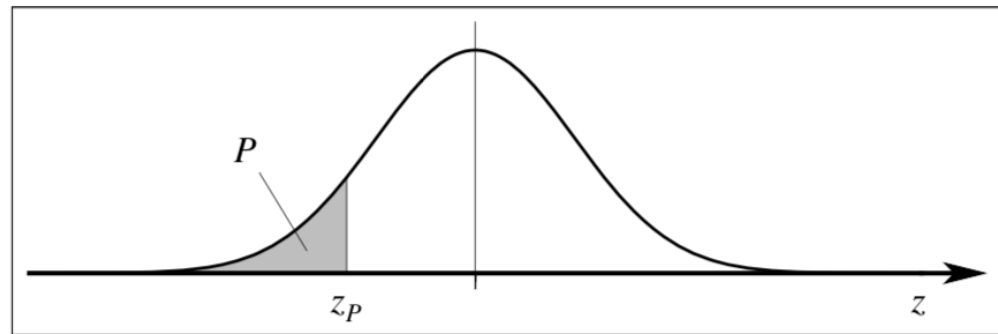


Figure A.1: Given a chosen z_p -value, the probability P (gray area under the curve from $-\infty$ to z_p) can be read from this table. Here, z_p is given in the format $-a.bc$, where $-a.b$ and $0.0c$ correspond to a unique row and column combination.

z_p	0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00
-3.4	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005
-3.2	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007
-3.1	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010
-3.0	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013
-2.9	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019
-2.8	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026
-2.7	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035
-2.6	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047
-2.5	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062
-2.4	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082
-2.3	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107
-2.2	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139
-2.1	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179
-2.0	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228
-1.9	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287
-1.8	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359
-1.7	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446
-1.6	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548
-1.5	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668
-1.4	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808
-1.3	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968
-1.2	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151
-1.1	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357
-1.0	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587
-0.9	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841
-0.8	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119
-0.7	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420
-0.6	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743
-0.5	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085
-0.4	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446
-0.3	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821
-0.2	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207
-0.1	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602
-0.0	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000

Table A.1: Normal cumulative distribution function. For $z > 0$ use $P(z) = 1 - P(-z)$.

TESTING OF HYPOTHESES

“The great tragedy of science – the slaying of a beautiful hypothesis by an ugly fact.”

Thomas Huxley, Biologist

Much of statistics is concerned with testing hypotheses for certain properties of entire populations based on sample data taken from these populations. There are numerous standard techniques used to perform these tests, and we will broadly group them into two sets:

1. Parametric tests
2. Nonparametric tests

This chapter will discuss the assumptions behind both kinds of tests and how they are performed.

4.1 The Null Hypothesis

At the core of all tests lies the concept of the “null hypothesis” (also known informally as the “boring hypothesis”). The null hypothesis, denoted H_0 , is stated and we will use our tests to see if we can reject it. For instance, if we want to test whether two rock types (i.e., two separate populations) have different densities, we obtain samples from each population and form the null hypothesis that they have equal densities (a boring result); we then test if we can reject H_0 . We will illustrate this approach with an example:

Example 4–1. It is claimed that the density of a particular sandstone unit is 2.35 g cm^{-3} . We are handed a sample of 50 specimens from an outcrop in the same area and decide to set the criteria that the sample comes from another lithological unit if the sample mean is less than 2.25 or larger than 2.45. In other words, our null hypothesis H_0 is $2.25 < \mu < 2.35$. This statement is a clear-cut criterion for accepting or rejecting the claim that the sample originates from the same unit, but it is not infallible. Since our decision will be based on a sample, there is the possibility that the sample mean \bar{x} may indeed satisfy $\bar{x} < 2.25$ or $\bar{x} > 2.45$ *even though* the population mean μ is 2.35. We will therefore want to know what the chances are that we could make a wrong decision and reject H_0 . Clearly, we must investigate what is the probability that $\bar{x} < 2.25$ or $\bar{x} > 2.45$ when μ in fact is 2.35. Here, $s = \sigma = 0.42$. This probability is given by the area under the two tails in Figure 4.1.

TESTING OF HYPOTHESES

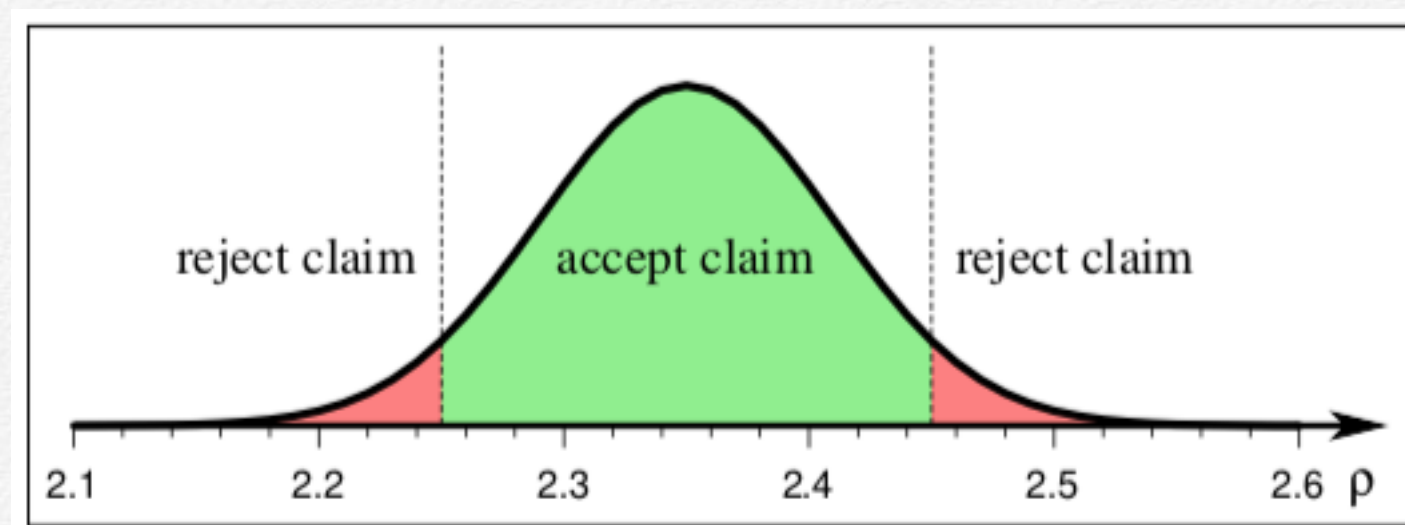
Since $n = 50 \gg 30$, we will treat our sample as of infinite size. We find the uncertainty in the sample mean to be

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{0.42}{\sqrt{50}} = 0.06. \quad (4.1)$$

We can now evaluate the normal scores of the two limits as

$$z_0 = \frac{2.25 - 2.35}{0.06} = -1.67, \quad (4.2)$$

Figure 4.1: Type I error: We reject the null hypothesis when the computed statistic falls in the tails and defer judgment if it falls in the central area.



$$z_1 = \frac{2.45 - 2.35}{0.06} = +1.67. \quad (4.3)$$

We find the area under each tail to be $\frac{\alpha}{2} = \frac{1}{2} \left[1 + \operatorname{erf} \left(-1.67 / \sqrt{2} \right) \right] = 0.0475$. Consequently, the probability of getting a sample mean that falls in the distal tails of the distribution is

$$\alpha = 2 \cdot 0.0475 = 0.095 \text{ or } 9.5\%. \quad (4.4)$$

This result means there is a 9.5% chance we will erroneously reject the hypothesis that $\mu = 2.35$ when it is in fact *true*. In statistics, we say we have committed a *type I error*¹.

TESTING OF HYPOTHESES

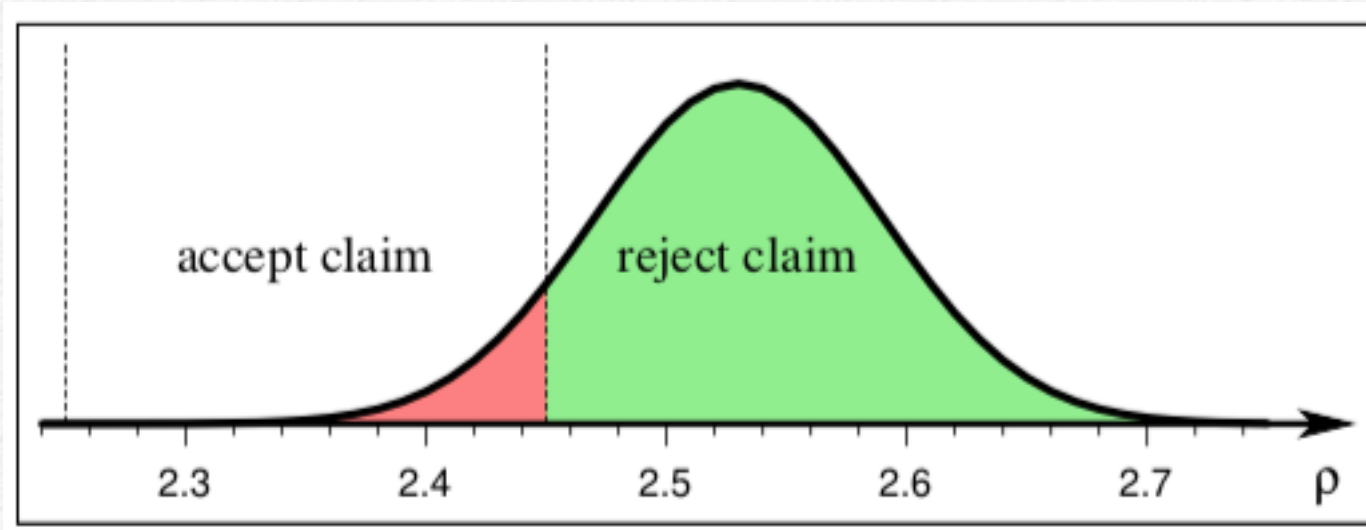


Figure 4.2: Type II error: Possibility of erroneously accepting an incorrect hypothesis.

Let us look at another possibility, where our test will fail to detect that μ is *not* equal to 2.35.

Example 4–2. Suppose, for the sake of argument, that the true mean $\mu = 2.53$. Then, the probability of getting a sample mean in the range 2.25 – 2.45, and hence erroneously accept the claim that $\mu = 2.35$, is now given by the tail area in Figure 4.2. As before, $s_{\bar{x}} = 0.06$ so the normal scores become

$$z_0 = \frac{2.25 - 2.53}{0.06} = -4.67, \quad z_1 = \frac{2.45 - 2.53}{0.06} = -1.333. \quad (4.5)$$

It follows that the probability is $\beta = \frac{1}{2} \left[\text{erf} \left(-1.333/\sqrt{2} \right) - \text{erf} \left(-4.67/\sqrt{2} \right) \right] = 0.092$ or 9.2%. This is the risk we run of accepting the incorrect hypothesis H_0 . We call this committing a *type II error*.

TESTING OF HYPOTHESES

Therefore, we recognize that there are several possible outcomes when testing a null hypothesis, as shown in Table 4.1. If the hypothesis is true, but is rejected, we have committed a Type I error, and the probability of doing so is designated α . In our example, α was 0.095. If our hypothesis is incorrect, but we still accept it, then we have committed a Type II error, and the probability of doing so is designated β . In our case, with $\mu = 2.53$, β was 0.092.

We saw in our example that the type II error probability *depended on the value of μ* . Since μ is typically not known, it is common to simply either reject H_0 or reserve judgment (i.e., never accept H_0). This way we avoid committing a type II error altogether, at the expense of never accepting H_0 . We call this a *significance test* and say that the results are *statistically significant* if we can reject H_0 . If not, the results are *not* statistically significant, and we attempt no further decisions. Of course, we may be wrong in rejecting H_0 but we can always state the likelihood of being wrong as α . Hence, in statistics we can only disprove hypotheses, but never prove them.

	Accept H_0	Reject H_0
H_0 is TRUE	Correct Decision	Type I Error
H_0 is FALSE	Type II Error	Correct Decision

Table 4.1: The four possible decision scenarios when testing an hypothesis. Of these, we always seek to avoid making a Type II error.

4.2 Parametric Tests

Parametric tests are used to make decisions based on parameters derived by assuming the data are approximately described by a probability distribution. The parameters typically used are properties of the distribution, such as the mean and standard deviations. Note that while we obtain our statistical parameters from the *sample*, our hypothesis testing applies to the *parent* population.

TESTING OF HYPOTHESES

4.2.1 What are the degrees of freedom?

When dealing with statistical tests we encounter the concept of “degrees of freedom” (usually given by the variable v) and it is often confusing to know what it is for various cases. We need to understand what v is in order to perform the tests. Typically, we will compute a *statistic* from our sample data. It may represent our best estimate of one *parameter* of the theoretical population distribution we are investigating via our limited sample. For instance, we may wish to compute a statistical quantity that requires us to use the mean of our data set of n points. While we will initially use $v = n$ in evaluating the mean (since all the points are independent), once we use the mean in *subsequent* calculations we must reduce v by one, since any individual data point can now be obtained from that mean and any $n - 1$ data points. In general, we say we lose one degree of freedom for each parameter we have estimated from the data. We will also see that certain transformations of the data, say, grouping the data into a set of k bins, will also affect v . While you may have had n data points to start with, after binning you now only have k items (the bin counts), so any testing involving these bins will have to consider a v that may only be $k - 1$ (you still lose one since the bin counts must sum to n). However, the degrees of freedom may be even less than $k - 1$ if any other parameters had to be estimated in order to facilitate the binning in the first place. In this Chapter, we will illuminate these situations using relevant examples.

4.2.2 Differences between sample means (equal variance)

We will often want to know if an observed difference between two sample means can be attributed to chance. We will again rely on the Student's t distribution. It is assumed that the two populations have the *same variance* but possibly *different means*.

We are interested in the distribution of $\bar{x}_1 - \bar{x}_2$, the difference in sample means. Of course, we only have one such estimate, but if the two samples are independent and random then the hypothetical difference distribution will be approximately normal, with mean $\mu_1 - \mu_2$ and standard deviation

$$\sigma_e = \sigma_p \sqrt{(1/n_1 + 1/n_2)}. \quad (4.6)$$

Here, σ_p is the *pooled* standard deviation,

$$\sigma_p = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}. \quad (4.7)$$

TESTING OF HYPOTHESES

We find the t -statistic by evaluating

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{\bar{x}_1 - \bar{x}_2}{s_e} \quad (4.8)$$

and test the hypothesis $H_0: \mu_1 = \mu_2$ based on the t -distribution for $v = n_1 + n_2 - 2$ degrees of freedom. Here, we have lost one degree of freedom for each sample mean we computed. For large n_1, n_2 , the t -distribution becomes very close to a normal distribution and we may instead use z -statistics based on

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \quad (4.9)$$

We will illustrate the two-sample t -test with an example.

Example 4–3. We have obtained random samples of magnetite-bearing rocks from two separate basalt outcrops. The measured magnetizations (in $\text{Am}^2\text{kg}^{-1}$) are

Outcrop 1: {87.4, 93.4, 96.8, 86.1, 96.4} $n_1 = 5$

Outcrop 2: {106.2, 102.2, 105.7, 93.4, 95.0, 97.0} $n_2 = 6$

We state our null hypothesis $H_0: \mu_1 = \mu_2$; the alternative hypothesis is of course $H_1: \mu_1 \neq \mu_2$. We decide to use a 95% confidence level, so the level of significance $\alpha = 0.05$. In this case, $v = 5 + 6 - 2 = 9$, and Table A.2 shows that the critical t value is 2.262. We will reject H_0 if our calculated t exceeds this critical value. Using the data, we find

$$\bar{x}_1 = 92.0 \text{ with } s_1 = 5.0, \quad (4.10)$$

$$\bar{x}_2 = 99.9 \text{ with } s_2 = 5.5. \quad (4.11)$$

Using (4.8) we obtain

$$t = \frac{99.9 - 92.0}{\sqrt{\frac{4.5^2 + 5.5 \cdot 5.5^2}{9} \left(\frac{1}{5} + \frac{1}{6} \right)}} = 2.5. \quad (4.12)$$

Since $t > 2.262$ we must reject H_0 . We conclude that the magnetizations at the two outcrops are not the same.

TESTING OF HYPOTHESES

A.2 Critical Values for the Student's t Distribution

One tail, α :	0.10	0.05	0.025	0.01	0.005
Two tails, α :	0.20	0.1	0.05	0.02	0.01
v					
1	3.078	6.314	12.71	31.82	63.66
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
∞	1.282	1.645	1.960	2.326	2.576

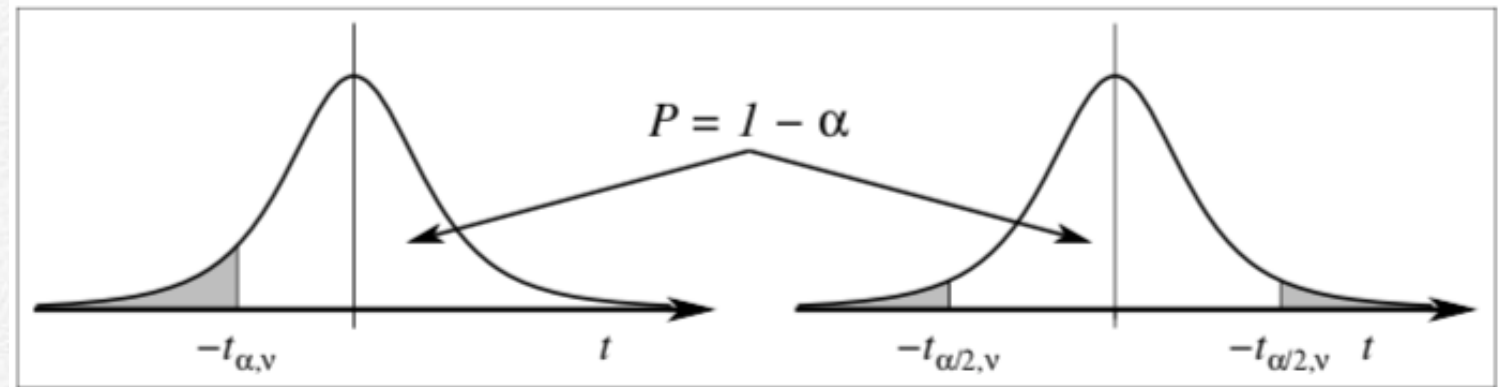


Figure A.2: Given a chosen confidence level, α (area of the tail(s)), and sample size, n , then the degrees of freedom, v , is $n - 1$. Find the corresponding row and column entries and read the critical $t_{\alpha, v}$ value (one-sided test, e.g., for $H_0: t < 0$) or $t_{\alpha/2, v}$ value (two-sided test for $H_0: t = 0$). The sign of t depends on which tail you are considering.

Table A.2: Critical values for the Student's t distribution.

TESTING OF HYPOTHESES

4.2.3 Differences between sample means (unequal variance)

In situations where the two populations *do not* have equal variance (which we will test for in Section 4.2.6) one can use *Welch's* *t*-test instead. It is similar to the regular two-sample *t*-test but now the observed *t* value is obtained directly from (4.9).

However, the estimation of the degrees of freedom differs and is given by

$$v \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 v_1} + \frac{s_2^4}{n_2^2 v_2}}. \quad (4.13)$$

Rounding *v* to the nearest integer allows you to look up critical *t* from Table A.2.

So far, we have put confidence limits on sample means and compared sample means to investigate whether two populations have different means. We will now turn our attention to inferences about the standard deviation.

4.2.4 Inferences about the standard deviation

The most popular way of estimating σ is to compute the sample standard deviation. When investigating properties of σ via *s* we will be using the “chi-square” statistic, given by

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}. \quad (4.14)$$

TESTING OF HYPOTHESES

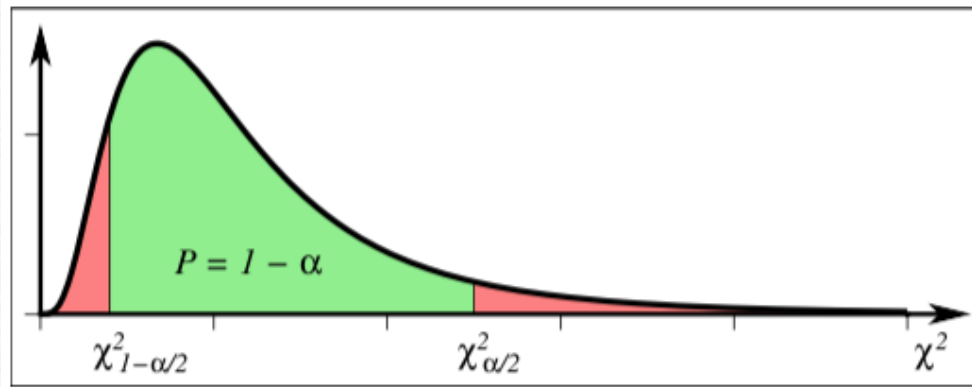


Figure 4.3: A typical chi-square probability density function, with mean value v and mode at $v - 2$.

The χ^2 distribution depends on the degrees of freedom, $v = n - 1$, and it is restricted to positive values because of the squared terms. It portrays how the sample variances would be distributed if we selected random samples of n items from a normal distribution with a standard deviation of σ . Figure 4.3 shows a typical χ^2 distribution. In the same way we used z_α and t_α , we now use χ^2_α as the value for which the area to the right of χ^2_α equals α . Because the distribution is not symmetrical, we must evaluate the critical values for $\alpha/2$ and $1 - \alpha/2$ separately. Furthermore, in the same way we placed confidence intervals on μ , we now use (4.14) to place confidence intervals on the variance, i.e.,

$$\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}} \quad (4.15)$$

or

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}, \quad (4.16)$$

which gives the α confidence interval on the variance. For large samples ($n > 30$) this can be simplified to

$$\frac{s}{1 + z_{\frac{\alpha}{2}}/\sqrt{2n}} < \sigma < \frac{s}{1 - z_{\frac{\alpha}{2}}/\sqrt{2n}}. \quad (4.17)$$

Note that the confidence interval is not symmetrical about the sample standard deviation.

TESTING OF HYPOTHESES

4.2.5 Testing a sample standard deviation

We might want to test whether our sample standard deviation s is equal to or different from a given population standard deviation σ . In such a case the null hypothesis becomes $H_0: s = \sigma$, with the alternative hypothesis $H_1: s \neq \sigma$. As usual, we select our level of significance to be $\alpha = 0.05$.

Example 4–4. We have 15 estimates of temperatures with $s = 1.3^\circ\text{C}$ and we want to know if s is significantly different from $\sigma = 1.5^\circ\text{C}$ based on past experience. From $\alpha = 0.05$ and $v = 14$ (since we lose one degree of freedom by first computing \bar{x} to obtain s), we find the critical χ^2 values from Table A.3 to be $\chi_{0.025}^2 = 26.119$ and $\chi_{0.975}^2 = 5.63$. Based on our sample statistic, we compute

$$\chi^2 = \frac{14 \cdot 1.3^2}{1.5^2} = 10.5. \quad (4.18)$$

We see that we cannot reject H_0 at the 0.05 significance level, so we simply reserve judgment. This was a two-sided test since we had to check that χ^2 did not land in either of the two tails.

For large samples ($n \geq 30$), the χ^2 critical values do not vary much with v and we may use the simpler statistic

$$z = \frac{s - \sigma}{\sigma / \sqrt{2n}} \quad (4.19)$$

and use the standard z -statistics, i.e., the $n = \infty$ entry in Table A.2.

TESTING OF HYPOTHESES

A.3 Critical Values for the χ^2 Distribution

Degrees of freedom, ν	Left tail ($1 - \alpha$)					Right tail (α)				
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.60
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	16.75
6	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80
16	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72
18	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16
19	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.034	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.260	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.886	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

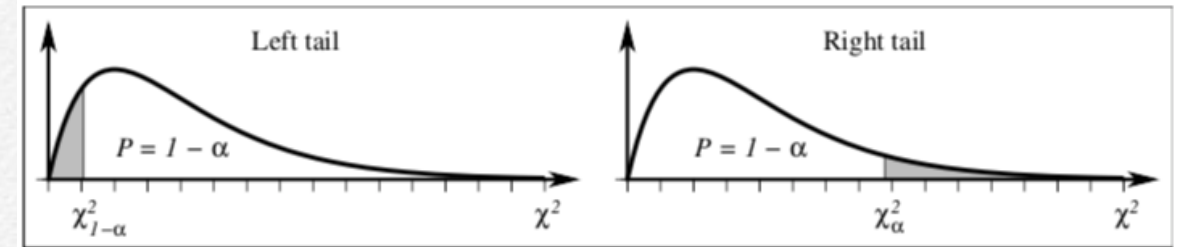


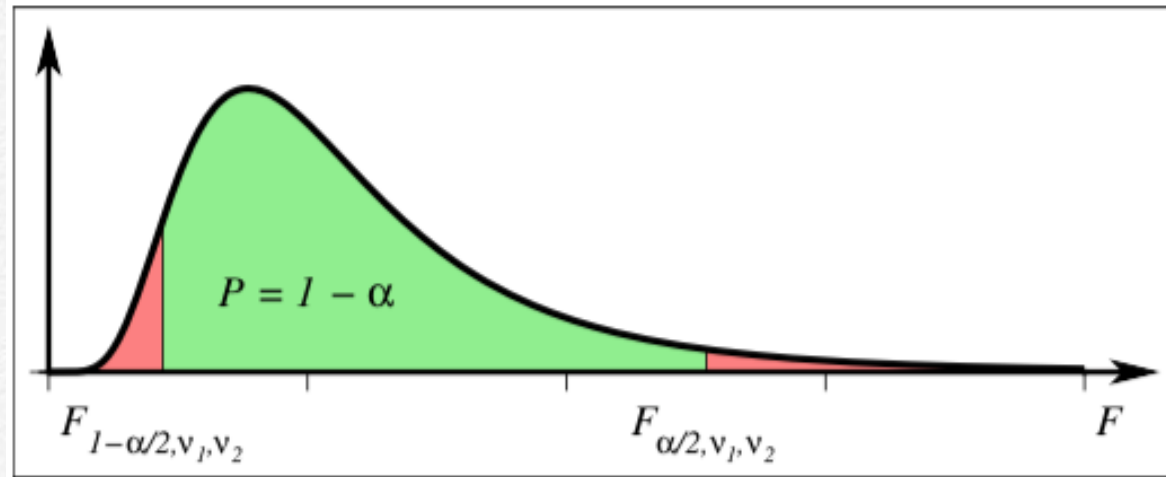
Figure A.3: Given α and degrees of freedom, $\nu = n - 1$, read critical $\chi^2_{1-\alpha, \nu}$ or $\chi^2_{\alpha, \nu}$ values.

Table A.3: Critical values for the χ^2 distribution.

TESTING OF HYPOTHESES

4.2.6 Comparing standard deviations from two samples

Figure 4.4: A typical F probability density function, for $\alpha = 0.10$, $v_1 = 20$ and $v_2 = 12$.



In the t -test for differences between two means (Section 4.2.2) we assumed that the standard deviations of the two samples were the same. Often this is not the case and one should first test whether this assumption is valid. We want to know whether the two variances are different or not. The statistic that is most appropriate for such tests is called the F -statistic, defined as the ratio

$$F = \begin{cases} s_1^2/s_2^2 & , s_1 > s_2 \\ s_2^2/s_1^2 & , s_2 > s_1 \end{cases} , \quad (4.20)$$

which typically is distributed as shown in Figure 4.4. For normal distributions, this variance ratio is a continuous distribution called the F -distribution. It depends on the two degrees of freedom, $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$. As before, we will reject the null hypothesis $H_0: \sigma_1 = \sigma_2$ at the α level of significance and (possibly) entertain the alternative $H_1: \sigma_1 \neq \sigma_2$ when our observed F -statistic exceeds the critical value $F_{\alpha/2}$. Note that because of the way (4.20) forces F to be equal to or larger than unity, any F -ratio that would have been located into the left tail is now instead mapped to the right tail. This is why we use $\alpha/2$ in determining the critical value as it is still a two-sided test.

TESTING OF HYPOTHESES

Example 4–5. In our case of rock magnetizations we assumed that the two standard deviations were approximately the same. Let us now show that this was actually justified. We find

$$F = \frac{5.5^2}{5.0^2} = 1.21. \quad (4.21)$$

From Table A.6 we find $F_{0.025}(v_1 = 5, v_2 = 4) = 9.36$. Hence we cannot reject H_0 and conclude that the difference in sample standard deviations is not statistically significant at the 95% level. In fact, s_1 would have to be over three times larger than s_2 ($\sqrt{9.36} = 3.06$ to be exact) before we would be able to reject H_0 .

4.2.7 Testing distribution shape: The χ^2 test

The next parametric test we shall be concerned with is the chi-squared test. It is a sample-based statistic using normal scores that are squared and summed up:

$$\chi^2 = \sum_{i=1}^n z_i^2 = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^2. \quad (4.22)$$

If we draw all possible samples of size n from a normal population and plotted a histogram of the resulting $\sum z^2$ they would approximate the χ^2 distribution mentioned earlier. The χ^2 test is used to compare the *shape* of our data distribution to a distribution of known shape (which is usually a normal distribution). The χ^2 test is most often used on data that have been categorized or *binned*. Assuming that our observations have been binned into k bins, the test statistic is found as

$$\chi^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}, \quad (4.23)$$

where O_j and E_j are the number of observed and expected values in the j 'th bin. At first glance you may think we have a typographical error in (4.23) — we do not. Note that this χ^2 still is nondimensional since we are using counts, even if the denominator is not squared. With counts, the probability that m out of n counts will fall in a given bin j is determined by the binomial distribution approximated by (3.3.1), with

TESTING OF HYPOTHESES

$$\bar{x} = E_j = np_j \quad (4.24)$$

and

$$s = \sqrt{np_j(1 - p_j)} \approx \sqrt{np_j} = \sqrt{E_j}. \quad (4.25)$$

Here, p_j is the probability that any value will fall in the j 'th bin. Plugging in for χ^2 , we find

$$\chi^2 = \sum_{j=1}^k \left(\frac{x_j - \bar{x}}{s} \right)^2 = \sum_{j=1}^k \left(\frac{O_j - E_j}{\sqrt{E_j}} \right)^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}. \quad (4.26)$$

as stated in (4.23). Since $\chi^2 = 0$ would mean a perfect match between observations and expectations we realize that this χ^2 test is *one-sided*, hence we will only check if our observed χ^2 statistic exceeds the critical $\chi^2_{\alpha, \nu}$ value.

Example 4–6. Consider the 48 measurements of salinity from Whitewater Bay in Florida (Table 4.2). We would like to know if these observations come from a normal distribution or not. The answer might have implications for models of mixing salt and fresh water in the bay.

The first step is to normalize the data into normal scores. We find $\bar{x} = 49.54$ and $s = 9.27$, and thus transfer all values via

$$z_i = \frac{x_i - 49.54}{9.27}. \quad (4.27)$$

We choose to bin the data into five bins whose boundaries were chosen so that the area under the curve for each bin is the same, i.e., 0.2. This choice ensures that the expected value will be the same for all bins (Figure 4.5).

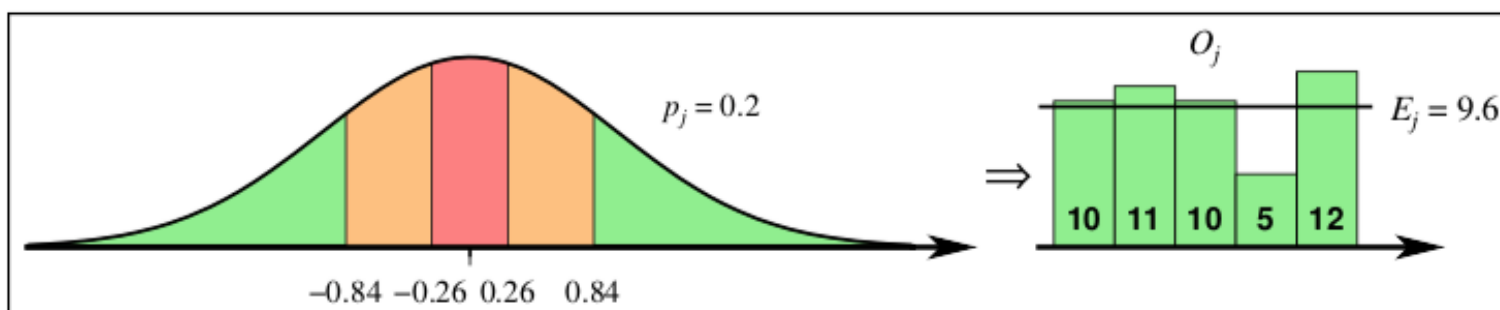


Figure 4.5: By desiring equal probabilities ($p = 0.2$) we obtain z -values for the bin boundaries that are not equidistant. This ensures that each bin has an acceptable number of observations and simplifies the calculation of the expected values.

TESTING OF HYPOTHESES

A.4.3 F table for 97.5% confidence level

$\downarrow v_2 \mid v_1 \rightarrow$	1	2	3	4	5	6	7	8	9	10	15	20	25	50
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	984.9	993.1	998.1	1008
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.43	39.45	39.46	39.48
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.25	14.17	14.12	14.01
4	12.22	10.65	9.979	9.605	9.364	9.197	9.074	8.980	8.905	8.844	8.657	8.560	8.501	8.381
5	10.01	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.428	6.329	6.268	6.144
6	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461	5.269	5.168	5.107	4.980
7	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.568	4.467	4.405	4.276
8	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	4.101	3.999	3.937	3.807
9	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964	3.769	3.667	3.604	3.472
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.522	3.419	3.355	3.221
11	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526	3.330	3.226	3.162	3.027
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	3.177	3.073	3.008	2.871
13	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250	3.053	2.948	2.882	2.744
14	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147	2.949	2.844	2.778	2.638
15	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060	2.862	2.756	2.689	2.549
16	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049	2.986	2.788	2.681	2.614	2.472
17	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061	2.985	2.922	2.723	2.616	2.548	2.405
18	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005	2.929	2.866	2.667	2.559	2.491	2.347
19	5.922	4.508	3.903	3.559	3.333	3.172	3.051	2.956	2.880	2.817	2.617	2.509	2.441	2.295
20	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774	2.573	2.464	2.396	2.249
21	5.827	4.420	3.819	3.475	3.250	3.090	2.969	2.874	2.798	2.735	2.534	2.425	2.356	2.208
22	5.786	4.383	3.783	3.440	3.215	3.055	2.934	2.839	2.763	2.700	2.498	2.389	2.320	2.171
23	5.750	4.349	3.750	3.408	3.183	3.023	2.902	2.808	2.731	2.668	2.466	2.357	2.287	2.137
24	5.717	4.319	3.721	3.379	3.155	2.995	2.874	2.779	2.703	2.640	2.437	2.327	2.257	2.107
25	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677	2.613	2.411	2.300	2.230	2.079
26	5.659	4.265	3.670	3.329	3.105	2.945	2.824	2.729	2.653	2.590	2.387	2.276	2.205	2.053
27	5.633	4.242	3.647	3.307	3.083	2.923	2.802	2.707	2.631	2.568	2.364	2.253	2.183	2.029
28	5.610	4.221	3.626	3.286	3.063	2.903	2.782	2.687	2.611	2.547	2.344	2.232	2.161	2.007
29	5.588	4.201	3.607	3.267	3.044	2.884	2.763	2.669	2.592	2.529	2.325	2.213	2.142	1.987
30	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511	2.307	2.195	2.124	1.968
40	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452	2.388	2.182	2.068	1.994	1.832
50	5.340	3.975	3.390	3.054	2.833	2.674	2.553	2.458	2.381	2.317	2.109	1.993	1.919	1.752
60	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270	2.061	1.944	1.869	1.699
70	5.247	3.890	3.309	2.975	2.754	2.595	2.474	2.379	2.302	2.237	2.028	1.910	1.833	1.660
80	5.218	3.864	3.284	2.950	2.730	2.571	2.450	2.355	2.277	2.213	2.003	1.884	1.807	1.632
90	5.196	3.844	3.265	2.932	2.711	2.552	2.432	2.336	2.259	2.194	1.983	1.864	1.787	1.610
100	5.179	3.828	3.250	2.917	2.696	2.537	2.417	2.321	2.244	2.179	1.968	1.849	1.770	1.592
110	5.164	3.815	3.237	2.904	2.684	2.525	2.405	2.309	2.232	2.167	1.955	1.836	1.757	1.577
120	5.152	3.805	3.227	2.894	2.674	2.515	2.395	2.299	2.222	2.157	1.945	1.825	1.746	1.565
130	5.142	3.796	3.218	2.886	2.666	2.507	2.386	2.291	2.213	2.148	1.936	1.816	1.737	1.555
140	5.134	3.788	3.211	2.879	2.658	2.500	2.379	2.284	2.206	2.141	1.929	1.808	1.729	1.546
150	5.126	3.781	3.204	2.872	2.652	2.494	2.373	2.278	2.200	2.135	1.922	1.801	1.722	1.538

Table A.6: Critical values for the F distribution, $\alpha = 0.025$. Note: v_1 , v_2 are the degrees of freedom for the numerator and denominator, respectively

TESTING OF HYPOTHESES

Using Table A.1, we find that the corresponding z-values for the intervals are $(-\infty, -0.84)$, $(-0.84, -0.26)$, $(-0.26, +0.26)$, $(0.26, 0.84)$, and $(0.84, \infty)$. Counting the values in Table 4.2 we find the observed number of samples for each of the five bins to be 10, 11, 10, 5, and 12. These are the observed O_j 's. The expected values E_j are all the same, i.e.,

$$E_j = \frac{n}{k} = \frac{48}{5} = 9.6. \quad (4.28)$$

Using (4.23) we find the observed value $\chi^2 = 3.04$.

The χ^2 distribution depends on ν , the degrees of freedom, which normally would be $\nu = k - 1 = 4$ in our case (we lose one since the bin counts must sum to n). However, we also used our observations to compute \bar{x} , then s , in order to *determine the bin boundaries*. These estimations further reduce ν by two, leaving just two degrees of freedom.

From the relevant Table A.3 we find critical χ^2 for $\nu = 2$ and $\alpha = 0.05$ to be 5.99. Since this is much larger than our computed value we conclude that we cannot, at the 0.05 level of confidence, reject the null hypothesis that the salinities were drawn from a normal distribution. We stress that while we used a normal distribution for comparison in this example, the E_j could have been derived from any other distribution we may wish to compare our data to.

Table 4.2: Standardized scores of salinity measurements from Whitewater Bay.

Sample Number	Original Sample	Standardized Sample	Sample Number	Original Sample	Standardized Sample
1	46.00	-0.38	25	35.00	-1.57
2	37.00	-1.35	26	49.00	-0.06
3	62.00	1.34	27	48.00	-0.17
4	59.00	1.02	28	39.00	-1.14
5	40.00	-1.03	29	36.00	-1.46
6	53.00	0.37	30	47.00	-0.27
7	58.00	0.91	31	59.00	1.02
8	49.00	-0.06	32	42.00	-0.81
9	60.00	1.13	33	61.00	1.24
10	56.00	0.70	34	67.00	1.88
11	58.00	0.91	35	53.00	0.37
12	46.00	-0.38	36	48.00	-0.17
13	47.00	-0.27	37	50.00	0.05
14	52.00	0.27	38	43.00	-0.71
15	51.00	0.16	39	44.00	-0.60
16	60.00	1.13	40	49.00	-0.06
17	46.00	-0.38	41	46.00	-0.38
18	36.00	-1.46	42	63.00	1.45
19	34.00	-1.68	43	53.00	0.37
20	51.00	0.16	44	40.00	-1.03
21	60.00	1.13	45	50.00	0.05
22	47.00	-0.27	46	78.00	3.07
23	40.00	-1.03	47	48.00	-0.17
24	40.00	-1.03	48	42.00	-0.81

TESTING OF HYPOTHESES

4.2.8 Test for a correlation coefficient

We recall that the conventional correlation coefficient was defined as

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (4.29)$$

Often, we need to test if an observed r is *significant*. In such tests, r is our sample-derived estimate of ρ , the actual correlation of the population pairs. The most useful null hypothesis is simply $H_0: \rho = 0$. It can be shown that the sampling distribution of correlations for a population that has zero correlation ($\rho = 0$) is a normal distribution with mean $\mu = 0$ and $\sigma = \sqrt{(1 - r^2)/(n - 2)}$. Hence, a t statistic can be calculated as

$$t = \frac{r - \mu}{\sigma} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = \frac{r\sqrt{n - 2}}{\sqrt{1 - r^2}}. \quad (4.30)$$

The degrees of freedom, v , is $n - 2$ since we needed to compute both \bar{x} and \bar{y} first to determine r .

Example 4–7. Suppose we roll a pair of dice, one red and one green (Table 4.3). Using (4.29) we obtain $r = 0.66$ which seems quite high, especially since there is no reason to believe a correlation should exist at all. Let us run a test to determine if the correlation is significant. Choosing $\alpha = 0.05$, Table A.2 shows that critical $t_{\alpha/2,3} = 3.182$. Applying (4.30) gives the observed $t = 1.52$, hence the correlation of 0.66 is most likely caused by random fluctuations found in small samples and (gratefully) we cannot reject H_0 .

How high would r have to be for us to find it significant and commit a type I error by rejecting the (true) null hypothesis? We can solve for r in

$$t_{\alpha/2, n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \Rightarrow 3.182^2 = \frac{3r^2}{1-r^2} \Rightarrow r = \pm 0.88. \quad (4.31)$$

Table 4.3: Result of tossing a pair of dice five times.

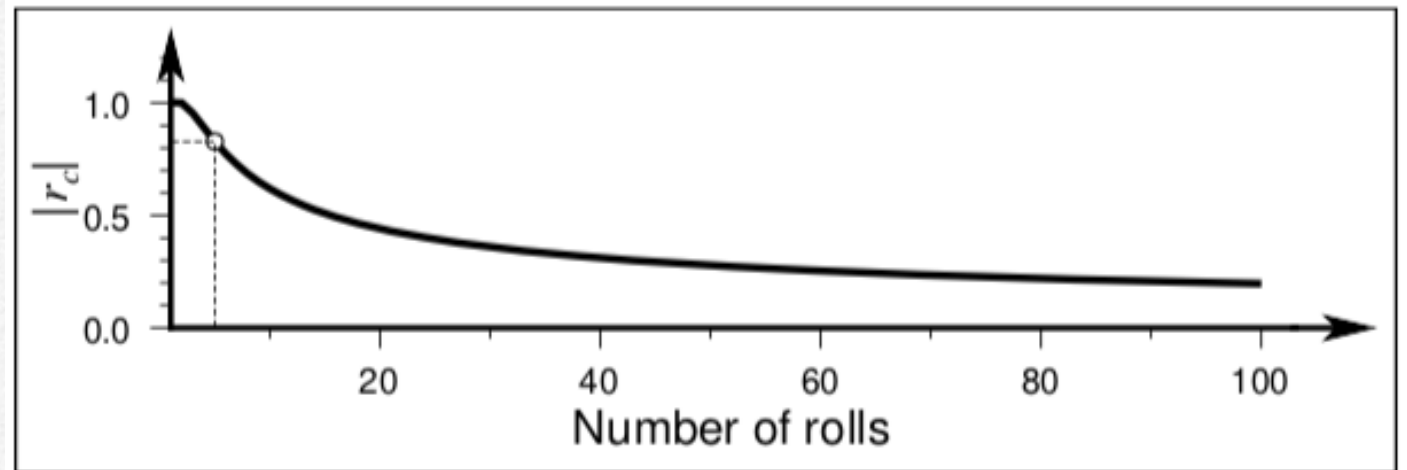
Red (x)	Green (y)
4	5
2	2
4	6
2	1
6	4

So, if r happens to equal or exceed ± 0.88 we would find ourselves awkwardly concluding that red and green dice give correlated pairs of values, but we would only make that mistake once in twenty tries on average.

TESTING OF HYPOTHESES

Going a bit further, we can use (4.30) to determine what the critical correlation, r_c , has to be for a given sample size (i.e., number of rolls) before we would suspect the dice have been tampered with. We simply solve the equation for r given a level of confidence and sample size. The results are displayed in Figure 4.6. High correlations are required when the number of rolls are relatively small.

Figure 4.6: The critical values of correlation, i.e., how large the observed correlation would have to be before we must reject the null hypothesis. The circle and dashed lines reflect the situation in Example 4.7.



4.2.9 Analysis of variance

We found earlier that we could use the Student's t -test to decide if two samples had different means. However, very often we are faced with the task of deciding whether observed differences among *more* than two sample means can be attributed to chance, or whether there are real differences among the means of the populations sampled. Statistical scientists developed an exploratory procedure to analyze such observations known as *ANOVA*, which stands for “ANalysis Of VAriance”. ANOVA usually comes in two flavors: One-way and two-way.

One-way ANOVA

In one-way ANOVA, the basic idea is to express the total variation of the data as a sum of two terms, each of which can be attributed to a specific *source*. The two sources of variation are:

1. Actual differences (different means) among the populations that the samples represent.
2. Chance or experimental error.

TESTING OF HYPOTHESES

The measure of the total variation that we shall use here is the *total sum of squares*, SST , which is simply

$$SST = \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{\bar{x}})^2, \quad (4.32)$$

where x_{ij} is the i 'th observation from the j 'th sample. Thus, we have k samples with n observations each and $\bar{\bar{x}}$ is the grand mean of all observations. Note that if we divided SST by $kn - 1$ we would obtain the variance of the combined data set. Let \bar{x}_j be the mean of the j 'th sample. We can then rewrite (4.32) as

$$\begin{aligned} SST &= \sum_j^k \sum_i^n [(\bar{x}_j - \bar{\bar{x}}) + (x_{ij} - \bar{x}_j)]^2 = \sum_j^k \sum_i^n [(\bar{x}_j - \bar{\bar{x}})^2 + 2(\bar{x}_j - \bar{\bar{x}})(x_{ij} - \bar{x}_j) + (x_{ij} - \bar{x}_j)^2] \\ &= n \sum_j^k (\bar{x}_j - \bar{\bar{x}})^2 + \sum_j^k \sum_i^n (x_{ij} - \bar{x}_j)^2 + 2 \sum_j^k \sum_i^n (\bar{x}_j x_{ij} - \bar{x}_j^2 - \bar{\bar{x}} x_{ij} + \bar{\bar{x}} \bar{x}_j) \\ &= n \sum_j^k (\bar{x}_j - \bar{\bar{x}})^2 + \sum_j^k \sum_i^n (x_{ij} - \bar{x}_j)^2 + 2 \sum_j^k \bar{x}_j n \bar{\bar{x}} - 2n \sum_j^k \bar{x}_j^2 - 2\bar{\bar{x}} \sum_j^k \sum_i^n x_{ij} + 2\bar{\bar{x}} n \sum_j^k \bar{x}_j \\ &= n \sum_j^k (\bar{x}_j - \bar{\bar{x}})^2 + \sum_j^k \sum_i^n (x_{ij} - \bar{x}_j)^2 + 2n \sum_j^k \bar{x}_j^2 - 2n \sum_j^k \bar{x}_j^2 - 2\bar{\bar{x}} kn \bar{\bar{x}} + 2\bar{\bar{x}} nk \bar{\bar{x}} \\ &= n \sum_j^k (\bar{x}_j - \bar{\bar{x}})^2 + \sum_j^k \sum_i^n (x_{ij} - \bar{x}_j)^2. \end{aligned} \quad (4.33)$$

Looking at the last two terms remaining we see that the first is a measure of the variation *among* the sample means. Similarly, the second term is a measure of the variation *within* the individual samples. It is customary to refer to the first term as the *treatment sum of squares*, $SS(Tr)$ and to the second term as the *error sum of squares*, SSE (Note: these names originated from agricultural experiments, i.e., “treatments” of different fertilizers.) Our overall goal with the ANOVA test is to investigate if the sample means are all statistically equal, hence the null hypothesis is

$$H_0 = \mu_1 = \mu_2 = \dots = \mu_k \text{ versus } H_1: \text{they are not all equal.}$$

TESTING OF HYPOTHESES

We usually state that $\mu_i = \mu + \varepsilon_i$, so $\sum \varepsilon_i \equiv 0$. Thus, the null hypothesis can also be written,

$$H_0 : \varepsilon_1 = \varepsilon_2 = \dots = \varepsilon_k = 0.$$

Let us introduce two assumptions that are critical to this test:

1. The populations we are sampling are approximately normal.
2. They all have the same variance, σ^2 .

If true, then we can look upon the k samples as if they came from the same (normal) population and therefore consider the variance of their means $s_{\bar{x}}^2$ as an estimate of $\sigma_{\bar{x}}^2$. Since $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ for infinite populations (recall the central limits theorem), we find that $ns_{\bar{x}}^2 = \sigma^2$. This is, upon examination, the same as $SS(Tr)/(k-1)$.

Since σ^2 is not known it must be estimated from the data. Furthermore, since the populations are assumed to be normal then any one of the sample variances could be used. To improve on the estimate we chose to take the mean of the sample variances,

$$\frac{s_1^2 + s_2^2 + \dots + s_k^2}{k}. \quad (4.34)$$

This turns out to equal the (normalized) second variance term, i.e., $SSE/(k(n-1))$. We now have two competing estimates of σ^2 for the population variance σ^2 . If the first estimate (based on variation among the sample means) is much larger than the second estimate (based on variations within the samples, i.e., variation due to chance) we should reject H_0 . This is simply the F -test comparing two variances. Consequently, we shall use the statistic

$$F = \frac{SS(Tr)/_{k-1}}{SSE/_{k(n-1)}}. \quad (4.35)$$

where the numerator is our estimate of σ^2 from variation among the \bar{x}_i while the denominator is our estimate of σ^2 from variation within the samples.

TESTING OF HYPOTHESES

The expected value of F is therefore unity. In practice, we construct an ANOVA table (Table 4.4).

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Treatments	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k - 1}$	$\frac{MS(Tr)}{MSE}$
Error	$k(n - 1)$	SSE	$MSE = \frac{SSE}{k(n - 1)}$	
Total	$kn - 1$	SST		

Table 4.4: Table used for a one-way analysis of variance (ANOVA).

We will apply this procedure to a small data set of four samples with six observations each (Table 4.5).

Example 4–8. The porosity of a sandstone was measured at four different locales and the observed values are listed in Table 4.5. At the 0.05 level of significance, is the porosity the same at all four locations? The null hypothesis becomes

$$H_0: \bar{\varphi}_1 = \bar{\varphi}_2 = \bar{\varphi}_3 = \bar{\varphi}_4. \quad (4.36)$$

We first find

$$\bar{x}_1 = 22.97, \bar{x}_2 = 22.25, \bar{x}_3 = 21.73, \bar{x}_4 = 19.83, \bar{\bar{x}} = 21.70 \quad (4.37)$$

The resulting ANOVA table is given as Table 4.6. From Table A.5 we find the critical F value for 3 and 20 degrees of freedom at the 5% level to be $F_{0.05,3,20} = 3.10$ (This is a one-sided test since we only consider if $F > 1$ or not). Since our observed F is much larger than the critical value we must reject our null hypothesis: The porosity at the four locations are not all the same.

Loc A	Loc B	Loc C	Loc D
23.1	21.7	21.9	19.8
22.8	23.0	21.3	20.4
23.2	22.4	21.6	19.3
23.4	21.1	20.2	18.5
23.6	21.9	21.6	19.1
21.7	23.4	23.8	21.9

Table 4.5: Four samples of sandstone porosity take from different locations.

Source of Variation	v	SS	MS	F
Treatments	3	32.35	10.78	10.78
Error	20	20.02	1.000	
Total	23	52.37		

Table 4.6: One-way ANOVA table based on the statistics from the porosity data given in Table 4.5.

TESTING OF HYPOTHESES

To simplify the calculations, we rewrite the expressions for the various sums of squares as

$$SST = \left\{ \sum_{j=1}^k \sum_{i=1}^n x_{ij}^2 \right\} - \frac{1}{kn} S^2, \quad (4.38)$$

$$SS(Tr) = \left\{ \frac{1}{n} \sum_{j=1}^k S_j^2 \right\} - \frac{1}{kn} S^2, \quad (4.39)$$

and

$$SSE = SST - SS(Tr). \quad (4.40)$$

Here, S_j is the sum of the values in the j 'th sample, and S is the total sum of all observations.

So far we assumed that each sample has the same number of observations. Instead, if there are n_j observations in the j 'th sample we get

$$SST = \left\{ \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}^2 \right\} - \frac{1}{N} S^2, \quad (4.41)$$

$$SS(Tr) = \left\{ \sum_{j=1}^k \frac{S_j^2}{n_j} \right\} - \frac{1}{N} S^2, \quad (4.42)$$

Where $N = \sum_{j=1}^k n_j$. Also, $v_{\text{treat}} = k - 1$ but $v_{\text{error}} = N - k$.

Two-way ANOVA

The previous ANOVA test was concerned only with the task of checking if the means of the populations were the same. The two-way ANOVA procedure extends the test to whether there also are variations *across* the populations. In other words, the population mean for j 'th treatment and i 'th *block* is expected to be

$$\mu_{ij} = \mu + \epsilon_j + \gamma_i. \quad (4.43)$$

TESTING OF HYPOTHESES

Thus, ε_j are the treatment effects that vary from sample to sample, and γ_i are called the *block effects* and vary within each sample. Examples might be porosity at four locations where the ε_j represent differences among the locations and γ_i may represent variations with depth across all samples. Again, we test

$$H_0: \varepsilon_1 = \varepsilon_2 = \dots = \varepsilon_k = 0,$$

but now we also consider the second null hypothesis

$$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_n = 0.$$

To do this we must obtain a quantity, similar to the treatment sum of squares, which measures the variation among the block means. If we let S_i be the total of all values in the i 'th block (e.g., depth) and substitute it for S_j , sum over i instead of j , and swap n and k , we find the *block sum of squares* via

$$SSB = \left\{ \frac{1}{k} \sum_{i=1}^n S_i^2 \right\} - \frac{1}{kn} S^2. \quad (4.44)$$

Hence, we compute SST and $SS(Tr)$ as before, SSB as just given, and SSE now becomes

$$SSE = SST - [SS(Tr) + SSB]. \quad (4.45)$$

Hence, Table 4.7 shows the extended ANOVA table for two-way analysis.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Treatments	$k - 1$	$SS(Tr)$	$MS(Tr) = \frac{SS(Tr)}{k - 1}$	$\frac{MS(Tr)}{MSE}$
Blocks	$n - 1$	SSB	$MSB = \frac{SSB}{n - 1}$	$\frac{MSB}{MSE}$
Error	$(k - 1)(n - 1)$	SSE	$MSE = \frac{SSE}{(k - 1)(n - 1)}$	
Total	$kn - 1$	SST		

Table 4.7: Two-way table for the analysis of variance (ANOVA).

TESTING OF HYPOTHESES

Example 4–9. We have measured the nickel concentration in a shale at four locations where we have obtained three observations at different depths in the unit. Our data are given in parts per million (ppm), with depth increasing downward in Table 4.8. We want to do a two-way ANOVA to see if the variations among the locations and among observations at the same depth (our “blocks”) are similar at the 95% level of confidence. We state

$$H_0: \varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = 0,$$

$$\gamma_1 = \gamma_2 = \gamma_3 = 0.$$

	Loc 1	Loc 2	Loc 3	Loc 4	S_i
Depth 1	71	44	50	67	232
Depth 2	92	51	64	81	288
Depth 3	89	85	72	86	332
S_j	252	180	186	234	852

Table 4.8: Four samples of nickel concentrations from different locations, sorted by depth.

Following the procedure, we compute the total sum of squares to be

$$\sum \sum x^2 = 63,414. \quad (4.46)$$

Combined with the sums in Table 4.8 we find

$$SST = 63,414 - \frac{1}{12} (852)^2 = 63,414 - 60,492 = 2922, \quad (4.47)$$

$$SS(Tr) = \frac{1}{3} [252^2 + 180^2 + 186^2 + 234^2] - 60,492 = 1260, \quad (4.48)$$

$$SSB = \frac{1}{4} [232^2 + 288^2 + 332^2] - 60,492 = 1256, \quad (4.49)$$

$$SSE = 2922 - [1260 + 1256] = 406. \quad (4.50)$$

We construct a two-way ANOVA table presented as Table 4.9.

Source of Variation	ν	SS	MS	F
Treatments	3	1260	420	6.21
Blocks	2	1256	628	9.28
Error	6	406	67.67	
Total	11	2922		

Table 4.9: Two-way ANOVA table resulting from the statistics of the nickel concentrations given in Table 4.8.

With the help of Table A.5 we reach the following decisions:

Treatments (locations): Critical value $F_{0.05,3,6} = 4.76$, so we reject the hypothesis that $\varepsilon_i = 0$.

Blocks (depth): Critical value $F_{0.05,2,6} = 5.14$, so again we reject the hypothesis that all $\gamma_j = 0$.

In other words, we conclude that the average nickel concentration is not the same at the four locations, and that it is not the same at all depths.