Appendix E. A Review of Statistics

The intention of these notes is to provide a few statistical tools which are not covered in the text on error analysis used in the introductory physics labs.[1] Nevertheless, we'll start with mean and standard deviation just so that we have no confusion on the nomenclature being used here. On the whole, this is intended as a brief review. The derivations of the results will not be included since they can be found in most statistical texts. The emphasis is on explaining the concept.

Table of Contents

There are three kinds of lies:
lies, damned lies, and statistics.
—Disraeli

---

[1] There are numerous books on statistics. A simple, good text is by John R. **Taylor**, *An Introduction to Error Analysis*, Mill Valley: University Science, 1982. It is used in the UCSD introductory physics labs. It may not be the best book, but you may have paid for it already. The recommendation is to read through his text carefully, one more time. Another good text is by John A. **Rice**, *Mathematical Statistics and Data Analysis*, 2nd. Ed., Belmont, Calif.: Duxbury Press, 1995. Rice is a Professor of Mathematics at UCSD who taught MATH 183.

A very old but extremely good and easy to follow book is by Philip R. **Bevington**, *Data Reduction and Error Analysis for the Physical Sciences*, New York: MaGraw-Hill, 1969. A text that you can follow on your own is George W. **Snedecor** and William G. Cochran, *Statistical Methods*, 7th ed., Ames: Iowa State University Press, 1980. A simple text on probability is by Frederick **Solomon**, *Probability and Stochastic Processes*, Prentice-Hall, 1987. If you are lost in the mathematics, a good conceptual explanation can be found in a nice little book by John L. Phillips, How to Think about Statistics, 6th Ed., New York: W.H. Freeman, 2000.

There are, of course, books written with engineers in mind. An easy to read and useful text is written by Robert M. **Bethea**, Benjamin S. Duran, and Thomas L. Boullion, *Statistical Methods for Engineers and Scientists*, 3rd Ed., New York: Marcel Dekker, 1995. For reference, there is the handbook edited by Harrison M. **Wadsworth**, *Handbook of Statistical Methods for Engineers and Scientists*, New York: McGraw-Hill, 1990.

Mean, standard deviation and confidence limits

We have a handful of beach sand and we want to know the average diameter of the grains. In this case and in other statistics problems, we are looking at a population containing *N* members and are interested in a certain property *x* which has a certain population distribution. Let's presume that the sand grains have a discrete size distribution. If there are *k* different classes (diameters of sand grains) and with each x_i occurring in the population f_i times, the probability of finding each class is

$$P_i = \frac{f_i}{N} \quad , \quad \text{with } i = 1,\ldots, k$$

The ***population*** **mean** and population **variance** are

$$\mu = \sum_{i=1}^{k} P_i x_i \qquad \text{and} \qquad \sigma^2 = \sum P_i (x_i - \mu)^2 \tag{1}$$

and σ is the population **standard deviation**. (To make typing easier, we'll skip the indexes of the summation sign where the omission would not cause confusion.)

When $f_i = 1$ for each $i = 1,\ldots, N$, we write

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad \text{and} \qquad \sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{2}$$

The variance is a kind of average of its own. In its case, it is the average of the sum of squared deviations from the mean.

It is unlikely that we are crazy enough to measure every single sand grain. Likewise, when we do an experiment, we usually only take a *small* sample such that $n \ll N$. So we can only calculate the ***sample*** mean (or estimate mean) and *sample* variance of the population as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

The estimated *sample standard deviation* of the population is *s*.[1]

We can also define the population **coefficient of variation** (CV) as σ/μ, and the sample estimate CV as $s/\bar{x}$. This quantity is often reported as a percentage and used in describing the amount of variation in a population (or sample). In a good number of problems, the mean and the standard deviation often tend to change together, and the value of CV is relatively constant.

There are questions that we would like to answer: How close is $\bar{x}$ to μ? Is the sample size large enough? Do we have to repeat the experiment? We need to define the normal distribution before we can answer these questions.

---

[1] There is a very subtle difference between the standard deviation of a sample versus the standard deviation of the population as estimated from sample data, *i.e.*, s here. We'll skip this detail since it won't severely affect the explanation of the topics that we want to cover.

You may also wonder why we use $(n - 1)$ as the divider, or in formal terms, the **degrees of freedom**. By definition, s^2 is the average variability of a distribution of sample means. By the time we want to calculate s^2, we should have known $\bar{x}$ already. So with one quantity computed, we have one less unknown and one less degree of freedom.

The **normal distribution** of a continuous (random) variable *x* is defined as

$$N_{\mu,\sigma^2} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{3}$$

which is completely determined by the population mean μ and variance σ^2. When we solve different problems, they will have different values of mean and variance, so it would be nice to construct a standard curve that applies to all cases. We thus define the **standard normal variate** (or deviate)

$$Z = \frac{x-\mu}{\sigma}$$

with the resulting **standard normal distribution**

$$N_{0,1} = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-Z^2}{2}\right] \tag{5}$$

which has a mean of zero and standard deviation equal to 1 no matter what *x* represents or what the population mean and variance are.

In other words, the use of the standard deviate (also called the *z* score) makes it possible to compare data of different distributions by putting them on a standard "scale." The mean of each scale is the common reference point for different distributions and the scale is the difference from the mean normalized by the standard deviation.

Now back to our question of how reliable is our sample mean as our estimate of the population mean. We of course do not expect $\bar{x}$ to be identical to μ. We usually do not even know where the population mean resides. The answer as to how close is $\bar{x}$ to μ depends on the standard error of $\bar{x}$, which is a measure of the variability of a distribution of sample means.

If we repeat our experiment (repeat sampling), we will obtain different (no matter how slight) values of $\bar{x}$, but we do expect these $\bar{x}$ from the repeated sampling to be normally distributed with a mean equal to μ. This is true even if the original population is not normally distributed when the sample size *n* is large enough. The standard deviation *of the sample mean* $\bar{x}$ is [1]

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{6}$$

But these results are really not good enough because we want to do our experiment only once and still be able to find a quantitative measure of how accurate $\bar{x}$ is.

We now use the property that $\bar{x}$ is normally distributed in repeated sampling with mean μ and standard deviation $\sigma/\sqrt{n}$. Also assuming that we know σ, we can calculate the probability

[1] With two sentences, we have skipped an important chapter in statistics and the very important *central limit theorem*. The derivation of the standard deviation of the mean (also called standard error) is not difficult. Ibid, Solomon, chapters 9 and 10, and Taylor, chapter 5.

that $\bar{x}$ is in error within a certain range.[1] We first need to know the probability of finding $\bar{x}$ which is given by the area under the normal distribution curve. In terms of the standard normal deviate (the common basis of statistical tables), the probability is the the cumulative area between ±Z (*i.e.*, $x = \mu \pm Z\sigma$) is

$$P(\mu - Z\sigma \le x \le \mu + Z\sigma) = 2\,\frac{1}{\sqrt{2\pi}}\int_0^Z \exp\left[\frac{-Z^2}{2}\right]dZ$$

Some of the numbers are summarized below. (Statistical tables usually provide only half the area, from 0 to Z.)

Table 1. The cumulative area (probability) under the normal distribution curve of x between the interval $x = \mu \pm Z\sigma$.

Z	Cumulative area under normal curve
0.674	50%
1	68.26%
1.96	95%
2	95.44%
2.58	99%
3	99.74%

Thus for a given measurement of $\bar{x}$, we can be 95% certain than $\bar{x}$ will lie between

$$\mu - 1.96(\sigma/\sqrt{n}) \le \bar{x} \le \mu + 1.96(\sigma/\sqrt{n}) \qquad (7)$$

A more practical interpretation is to rewrite eq. (7) as the 95% **confidence limits** (or intervals) for μ

$$\bar{x} - 1.96(\sigma/\sqrt{n}) \le \mu \le \bar{x} + 1.96(\sigma/\sqrt{n}) \qquad (8)$$

The other (pessimistic) way of looking at it is that we have an unlucky 5% chance (1 in 20) that from our one experiment, $\bar{x}$ does not lie within the interval $\mu \pm 1.96\sigma/\sqrt{n}$. We can of course choose other levels of confidence limits.

The chance of enclosing the population mean μ increases if we choose to have a large interval. We also say that we have a "high level of confidence" in such a case. Of course we can include the population mean for sure if we choose the entire possible range of the variable x, but it would not be a useful answer. We usually draw the line at using the 95% confidence limits.

❸ ④ ❽ ❶ ⑨ ❿ ② ❺ ⑦

[1] How can we know σ when we don't know what μ is? We have to assume that we know σ from past data of similar populations. Later, the use of the *t* distribution avoids this problem.

The *t* distribution

In most applications, the value of σ is not known. To describe the variability of the measurement in the sample mean $\bar{x}$, we use the sample *standard error* [1]

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{6a}$$

We also define the test statistic, *t*, as analogous to the standard normal deviate Z, with (n – 1) degrees of freedom (because we already have used the sample mean once to calculate the sample standard deviation):

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad , \quad \text{with} \quad df = n - 1$$

The test statistic measures the discrepancy between the sample mean and the population mean. The discrepancy is normalized by the sample standard error. Without going into the theoretical details, let's just state that the *t* distribution accounts for sample size. For this reason, we always use the *t* statistic rather than the *Z* standard deviate. Very loosely speaking, the larger the sample size, the less "forgiving" is the distribution on the discrepancy. [2]

Confidence limits based on the *t* distribution

We can now approximate the confidence limit by replacing σ and Z in eq. (8). Using the 95% confidence limit as example, we can write

$$\bar{x} - t_{0.05}(s/\sqrt{n}) \leq \mu \leq \bar{x} + t_{0.05}(s/\sqrt{n}) \tag{8a}$$

The tabulation of the *t* distribution is different from the standard normal distribution. As opposed to reporting the area (probability) under the curve between μ and $\mu + Z\sigma$ (or simply between 0 and Z on the standardized curve), the *t* distribution table uses the area in *both tails outside* the interval $\bar{x} \pm ts_{\bar{x}}$.[3] The idea is that this value represents the probability $P(\bar{x} \geq \mu \pm ts_{\bar{x}})$. This is why for a 95% confidence level, we use the notation $t_{0.05}$.

When we read the *t* distribution table (typically in the Appendix of a statistics text) for a given value of *t* and df, the probability is referred to as "*probability of a larger value*." Say at *t* = 2.262 and df = 9, the *t* value falls under the 0.05 or 5% column. What it means is that there is a 5% chance that the discrepancy between $\bar{x}$ and μ (absolute value $|\bar{x} - \mu|$) is larger than 2.262 times the sample standard deviation. In other words, there is a 5% chance that our

---

[1] Properly speaking, the notation is the "sample standard deviation of the sample mean." Yes, it is very confusing. So most people just call it the standard error (SE).

[2] As engineers, we may think of *t* as the discrepancy as measured by the "number of sample standard error units." The *t* distribution is normally distributed but it depends on the sample size, using df = *n* – 1. The table of *t* distribution ignores the sign of *t*. The derivation of the distribution can be found in, for example, Rice, Chapter 6.

You may also find the reference to **Student's** *t* distribution. "Student" was the pseudonym used by the British statistician William S. Gossett in 1908. Gossett was working with the Guinness Brewery in Dublin when he thought of the idea and did the analysis. (See if you can remember this tidbit next time you have a Guinness.) Most texts now have dropped "Student" from the name of the distribution.

[3] Hence also the name "two-tailed tests."

measured sample mean is such that $\bar{x} \geq |\mu \pm 2.262 s_{\bar{x}}|$. Of course, there is a 95% chance that $\bar{x}$ is within the interval $|\mu \pm 2.262 s_{\bar{x}}|$. If we choose a smaller value of *t*, say *t* = 0.883 under the 0.4 column, we now have a 40% chance to find $\bar{x}$ outside the smaller interval, $|\mu \pm 0.883 s_{\bar{x}}|$.

Example 1

Consider the weights of 11 people: 148, 154, 158, 160, 161, 162, 166, 170, 182, 195, and 236 lb. With our calculator, we can get with n = 11, $\bar{x}$ = 172 lb, s = 24.95 lb, $s_{\bar{x}}$ = 24.95/√11 = 7.52 lb.

With df = 11 – 1 = 10 and looking up a t distribution table, we find $t_{0.05}$ = 2.228. Thus $ts_{\bar{x}}$ = (2.228)(7.52) = 16.75 lb and the 95% confidence interval ($\bar{x} \pm t_{0.05} s_{\bar{x}}$) extends from 155.25 to 188.75 (172 ± 16.75) lb.

A claim that μ (whatever it is) lies between 155.25 and 188.75 lb is likely to be correct with a 19-in-20 chance. Since 16.75/172 = 0.097 ≈ 0.1, we can also say that the estimate $\bar{x}$ = 172 is 95% chance to be correct to within ±10% of μ.

Tests of hypotheses using the *t* distribution

A test usually involves setting up a hypothesis[1] and comparing an expected mean with an observed average that contains the inherent variability within the data. The discrepancy is measured by the "number" of sample standard error "units," *t*. We provide the definitions for three common hypotheses and their corresponding definitions for *t*.

A. From one set of data, we assert that the sample mean is the same as the expected mean μ_o. The hypothesis is

$$H_o: \mu = \mu_o$$

and the resulting test statistic is

---

[1] The term is properly called the **null hypothesis** H_o in statistics. If we have to compare two quantities, the nominal practice is to argue (in a civilized way) that there is *no* difference (and hence the null). Under most circumstances, we try to *disprove* that there is no difference. One good example is doing least-squares fits. The null hypothesis, as will be shown later, is that the slope is zero, and which of course we'd like very much to be false.

Like virtually all statistical problems, we never know with absolute certainty what the "answer" is, so we always have to make our statement in probabilistic terms. The null hypothesis is phrased such that if it is true (*i.e.*, can be accepted within a given probability limit), there is no real change or difference. And H_o is false if the difference is significant enough that we can reject the hypothesis. If we cannot reject the null hypothesis, it does not mean that it is really true; it just means that the hypothesis is tenable.

In many problems, we also propose an **alternative hypothesis** H_1. For example, if H_o: $\mu = \mu_o$, we may propose H_1: $\mu < \mu_o$, *i.e.*, saying that if it is not true that $\mu = \mu_o$, then it may very well be that $\mu < \mu_o$.

$$t = \frac{\bar{x} - \mu_o}{s_{\bar{x}}}, \text{ with } df = n - 1$$

B. In comparing two independent sets of data, we assert that the means of the two sets of samples are the same.[1] The sample sizes of the two data sets are n_1 and n_2 respectively. The hypothesis is

$$H_o: \mu_1 = \mu_2$$

and the resulting test statistic (also called **unpaired-t statistic**) using sample means is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}}}, \quad \text{with} \quad df = n_1 + n_2 - 2$$

where now

$$s_{\bar{x}} = \bar{s}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and

$$\bar{s} = \sqrt{\frac{\sum(\bar{x}_1 - x_1)^2 + \sum(\bar{x}_2 - x_2)^2}{n_1 + n_2 - 2}}$$

C. In comparing two independent sets of data, the samples are *paired* (of course, $n_1 = n_2 = n$), and we assert that the two samples are the same. The hypothesis, which really says that there is no difference between each pair, is

$$H_o: \mu_1 - \mu_2 = 0$$

The test statistic is based on the differences, let's say x, between each pair of samples. Thus we test whether the mean of the measured differences $\bar{x}$ is the same as the expected difference, which is zero. Hence, the test statistic (also called **paired-t statistic**) is

$$t = \frac{\bar{x} - 0}{s_{\bar{x}}}, \quad \text{with} \quad df = n - 1$$

where the values of $\bar{x}$ and $s_{\bar{x}}$ are based on the differences of the two data sets.

We now go through several examples. *Read the wording slowly and carefully.* The most difficult part of these calculations is the thinking process. You should also make a copy of the *t* distribution, which can be found in the appendix of statistics textbooks.

---

[1] For more detailed explanation, see for example, Rice, chapter 11. There are also nonparametric methods such as the **Wilcoxon** rank sum and signed rank tests.

Example 2

First, we will focus on the first hypothesis: testing a sample mean. Say, a dairy farmer fed his cattle with some genetically engineered growth hormone. Later, he wants to know something about the weight gains of his cattle population. He takes a small sample of 10 cattle and measures the weight gain of each.

Sample a. The 10 weight gain measurements are 33, 53, 34, 29, 39, 57, 12, 24, 39, and 36 lb.

Take out your calculator and show that with n = 10, we can find $\bar{x}$ = 35.6 lb, s = 13 lb and $s_{\bar{X}}$ = 4.11 lb.

With df = 9 and $t_{0.05}$ = 2.262 (table lookup), $ts_{\bar{X}}$ = 9.3 and the 95% confidence limits for the population mean is between 26.3 and 44.9. In other words, there is only a 1-in-20 (or 5%) chance that the population mean does not fall within 26.3 and 44.9.

Now if we are being told that the average weight gain of the cattle population is 30 lb (*i.e.*, H_o: μ = 30, or μ_o = 30), we want to know if the sample mean supports this hypothesis at the 95% confidence level.

The *t* value is (35.6 – 30)/4.11 = 1.36 which is less than $t_{0.05}$ = 2.262. Repeating what we have said earlier, the discrepancy between $\bar{x}$ and μ_o has to be as much as 2.262 "standard error units" before there is a 5% chance that our sample mean falls outside the 95% interval (26.3 to 44.9). Since the discrepancy is now less (only 1.36), we cannot reject the hypothesis that the expected mean of the population weight gain might be 30 lb in the 95% *confidence* interval. Or "jargon-wise," at the 5% *significance* level.[1]

Another interpretation of the analysis is that if we refer back to the *t* distribution table at df = 9, a *t* value of 1.36 is roughly at the 0.2 probability level. Thus there is a 20% chance that the discrepancy between $\bar{x}$ and μ_o is larger than 1.36 "standard deviation units." In other words, there is a 20% chance that the expected mean lies outside the (smaller) interval $|\bar{x} \pm t_{0.2}s_{\bar{X}}|$ or |35.6 ± (1.36)(4.11)| which is 30.0 to 41.1. (Compare this with the 95% confidence interval.)

Another way to interpret the result is that if we reject the hypothesis that the expected mean is 30 lb, we have a 20% chance of making a mistake.

Sample b. The 10 weight gain measurements are 17, 22, 20, 19, 3, 21, 25, 40, 21, and 36 lb.

---

[1] It is always odd to see that the **significance level** is a small number the first time. Why is the "significance level" the opposite of the "confidence level"?

By convention, the *significance refers to the probability that a null hypothesis can be rejected.* So this leads to the seemingly perverse thinking that when the significance is a *small* value, we have a *high* confidence that we can reject the null hypothesis. Since the null hypothesis is usually constructed to assume no difference between two quantities, we say that the difference is statistically significant if we can reject the hypothesis. To avoid confusion, we rarely use the terms confidence and significance in the same sentence.

In another view, we try to show that the difference between the quantities cannot occur purely by chance. That is, the random chance that there is a significant difference better be small.

This time, we should find $\bar{x}$ = 19.1 lb, s = 10.6 lb and $s_{\bar{x}}$ = 3.35 lb.

With df = 9 and $t_{0.05}$ = 2.262 (table lookup), $ts_{\bar{x}}$ = 7.6 and the 95% confidence limits for the population mean is between 11.5 and 26.7. Again, we presume that there is only a 1-in-20 chance that the population mean does not fall within 11.5 and 26.7.

We repeat the same game that presumes the average weight gain of the cattle population to be 30 lb (*i.e.*, the μ_o), and we want to know if the measured sample mean supports this hypothesis at the 95% confidence level.

The *t* value is (19.1 – 30)/3.35 = –3.25, or 3.25 ignoring the sign. The value is larger than $t_{0.05}$ = 2.262. With the large discrepancy (3.25 "standard deviation units"), there is more than a 5% chance that our sample mean of 19.1 falls outside the 95% interval (11.5 to 26.7). So we can reject the hypothesis that the expected mean is 30. Another way we can look at the result is that if it were true that $\bar{x} = \mu_o$, we shouldn't (on average) have a discrepancy as large as $3.25s_{\bar{x}}$. Thus it is unlikely that $\bar{x} = \mu_o$ and we can reject this claim at least at the 95% confidence level.

Now there are two possibilities which we cannot distinguish yet. One is that the expected mean of 30 lb that somebody told us was wrong. The second possibility is that someone actually had weighed the hundreds of cattle and there was something wrong with our experiment.

Also, at df = 9, a *t* value of 3.25 is roughly at the 0.01 probability.[1] There is only a 1% chance that the expected mean lies outside the interval |19.1 ± (3.25)(3.35)| or 8.21 to 29.99. And if we reject the hypothesis that the expected mean is 30 lb, we only have a 1% chance of making a mistake.

The second sample shows that we could still make the wrong statement if we just calculate the 95% confidence interval.

Example 3

In a production facility, we have two production lines making some widgets. We want to know if the two assembly lines are making the same product. For this test, "sameness" is determined by weight. We take some samples off of each line:

Line 1	Line 2
65.0 kg	64.0 kg
64.5	69.0
74.5	61.5
64.0	69.0
75.0	67.5
74.0	
67.0	

The hypothesis that we are testing is whether the means of each of these samples are equal (unpaired *t* test). First, we need means and variances for each assembly line:

[1] "Jargon-wise," you may say that the *p*-value is 0.01 or less and hence we can conclude that the measured mean and the expected mean are *significantly* different.

Line 1	Line 2
$\bar{x}_1 = 69.1$	$\bar{x}_2 = 66.2$
$n = 7$	$n = 5$
$s_{x1} = 5.11$	$s_{x2} = 3.32$
$s_{x1}^2 = 26.1$	$s_{x2}^2 = 11.0$

For this example, the degrees of freedom are $n_1 + n_2 - 2 = 10$. Next we need to calculate $\bar{s}$ and then $s_{\bar{x}}$:

$$\bar{s} = \sqrt{\frac{\Sigma(\bar{x}_1 - x_1)^2 + \Sigma(\bar{x}_2 - x_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(n_1 - 1)\, s_{x1}^2 + (n_2 - 1)\, s_{x2}^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(7-1)(26.1) + (5-1)(11.0)}{7 + 5 - 2}} = 4.47$$

$$s_{\bar{x}} = 4.47 \sqrt{\frac{1}{7} + \frac{1}{5}} = 2.62$$

Finally, we calculate the unpaired *t* statistic:

$$t = \frac{69.1 - 66.2}{2.62} = 1.11$$

When we look on the table for the *t* value for 10 degrees of freedom, we see that 2.228 is the value for the 95% interval (5% significance). Since our *t* value is less than that, we can not reject the hypothesis that the two production lines have the same mean. Another way of looking at it is that 1.11 falls between 20 and 30% (in the *t* distribution table). That would mean that if we reject the hypothesis that they are the same, we have a 20 to 30% chance of being right. Not very good odds of getting something right!

Example 4

Now, we'll consider the use of the paired t test. Seven automotive catalytic converters are tested for their activities. Afterward, they are exposed to leaded gasoline for a period of time. Their activities are tested a second time and we want to know if the exposure to leaded gasoline may have altered the catalytic converters. The data are:

Sample	Analysis 1	Analysis 2	Difference
1	15.1	14.6	– 0.5
2	14.7	14.2	– 0.5
3	15.2	15.0	– 0.2
4	15.3	15.3	0.0
5	14.9	14.0	– 0.9
6	14.7	14.6	– 0.1
7	15.1	14.5	– 0.6

The average difference is – 0.4. First we need to know the sample standard deviation, s, and the standard error of the sample mean, $s_{\bar{x}}$. The result is s = 0.316 and $s_{\bar{x}}$ = 0.119.

The hypothesis that we are testing is whether $\mu_1 - \mu_2 = 0$, or if $\bar{x} = 0$. For 6 degrees of freedom, we calculate the *t* statistic as

$$t = \frac{-0.4 - 0}{0.119} = -3.36$$

We look up the table for *t* values, and see that for 6 degrees of freedom, the 95% significance value is 2.45. Since the value that we calculated is greater than this (ignoring sign) there is better than a 95% probability that the two analyses are not the same, *i.e.*, the hypothesis that they are the same may be rejected with a likelihood of less than 5% error.

In each problem, we have a hypothesis, the null hypothesis. If we make the mistake of rejecting the null hypothesis while it is actually true, we make what statisticians called a **type I** (or α) **error**. On the other hand, if we stick with the null hypothesis even if it is not true, the mistake is called a **type II** (or ß) **error**.

In most examples, we pick the 5% probability (significance level) as the criterion. When we reject the null hypothesis if the test statistic is larger than the value according to the *t* distribution, we have a 5% risk of making a type I error. Of course, we can use a much smaller significance level, say, 0.1%, in which case we need a much larger value of test statistic to reject the hypothesis. The risk we now face is that we can mistakenly accept the hypothesis when we should have rejected it—making a type II error. In a way, choosing the 5% significance level is a compromise between making the two possible kind of errors.

Regression of two variables

Regression as a functional relation of correlated variables. For two variables x and y, we assume that they are related by the linear relation:

$$y = \alpha + \beta(x - \bar{x}) \tag{9}$$

which we consider as the *population* regression line. Thus for any given values of x, the corresponding $y = y(x)$ is independently and normally distributed with a population mean $\mu_y(x)$ that lies on the straight line and the deviations from the regression line have a constant variance σ_y^2. [1]

We now try to describe the relation with our statistical model, *i.e.*, to predict the values of y, with

$$\begin{aligned} \hat{y} &= \bar{y} + b(x - \bar{x}) \\ &= (\bar{y} - b\bar{x}) + bx \end{aligned} \tag{10}$$

which is the form preferred in many statistical packages. We also call the model as *the linear regression of y on x*. For us, the more common arrangement is

$$\hat{y} = a + bx$$

where $a = (\bar{y} - b\bar{x})$.

In a way, the model is described by

$$\hat{y} = \bar{y} + b(x - \bar{x}) \quad = \alpha + \beta(x - \bar{x}) + \varepsilon$$

where ε is a random error that is normally distributed with mean zero and constant variance σ_y^2.

Of course, when we do an experiment, we collect a small number of discrete pairs of data points. For a set of data consisting of n pairs of measurements (x_i, y_i), we can evaluate the coefficients for the regression line by minimizing the sum of squared deviation of the data from the prediction:

$$\min\left[\Sigma(y_i - \hat{y}_i)^2\right] = \min\left[\Sigma(y_i - a - bx_i)^2\right]$$

This is why we also call the procedure a **least-squares fit**. We just state the results: [2]

$$a = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i}{\Delta}$$

---

[1] A good schematic illustration is given on p. 154 of Snedecor and Cochran. An implicit assumption is that we are certain about the values of x, which, while not always true, it is generally true that we are much more sure about the value of x than of y.

As for the term "regression," it came from the geneticist Sir Francis Galton (1822-1911) who found that offspring tended to be smaller than their parents if their parents were larger than the average population, while the offspring tended to be larger than their parents if their parents were smaller than average. He called the statistical observation "regression towards mediocrity."

[2] The derivations are in, for example, Taylor or Bevington.

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\Delta}$$

where

$$\Delta = n\sum x_i^2 - (\sum x_i)^2$$

All the summations are meant to go from i = 1 to n.

The population variance of the fit is σ_y^2, and we calculate instead the *sample* (estimate) variance

$$s_{y_i}^2 = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2} \tag{11}$$

which is the residual sum of squares divided by the degrees of freedom. We also have added the subscript *i* to the nomenclature to reflect the fact that we are using discrete data points (x_i, y_i). We can easily show that

$$s_{y_i}^2 = \frac{\sum(y_i - \bar{y})^2 - b^2\sum(x_i - \bar{x})^2}{n - 2}$$

The sample standard deviation of the fit is s_{y_i}.

We can also derive the sample variance of the slope *b*: [1]

$$s_b^2 = \frac{s_{y_i}^2}{\sum(x_i - \bar{x})^2} \tag{12}$$

or using the notations that we defined above

$$s_b^2 = \frac{ns_{y_i}^2}{\Delta}$$

With the sample estimate of the standard error of b, s_b, we can evaluate the test statistic

$$t = \frac{b - \beta}{s_b} \tag{13}$$

If we make the hypothesis that $\beta = 0$ (or really H_o: $b = 0$), we can use the *t* distribution calculation to test if we can reject the possibility that *y* is not linearly correlated to *x*. (Our biased wish, in most cases, is that *y* is indeed linearly correlated to *x* with little probability of random error.)

From the *t* distribution table and a given degree of freedom, we can look up the probability level that is associated with the value of the test statistic. We may come across the "statement" [2]

$$p < 0.01,$$

---

[1] See, for example, chapter 6 in Bevington or chapter 9 in Snedecor and Cochran.

[2] Strictly speaking the **p value** is the smallest level of significance (see Example 2) with which we could reject the null hypothesis. Commonly, we simply think of the p value as the probability that we may have mistakenly rejected H_o even if it is true (a type I error). So the smaller the p value, the more boldly (certain) we can make the rejection.

which is commonly used by biologists and social scientists, often without explanation. (Of course, the value 0.01 may change.) Commonly, they just state the probability with the results of a least-square fit. What they, in essence, are saying: "We can reject the hypothesis that there is no association between *x* and *y*. The evidence is so strong that there is less than a 1% chance that the fit is a random fluke."

For us, an important question is finding the variance of the predicted value $\hat{y}_i$ for a given value of x_i. The sample variance of the population regression line for the prediction of $\hat{y}_i$ is[1]

$$s^2_{\hat{y}} = s^2_{y_i}\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right)$$

At $x_i = 0$, $\hat{y}_i = a$, the intercept, and $s^2_{\hat{y}} = s^2_a$ and we can easily show that the sample variance of the intercept is

$$s^2_a = \frac{s^2_{y_i}\sum x_i^2}{\Delta}$$

As opposed to just reporting the standard errors in the slope and the intercept, we can report, for example, the 95% **confidence intervals** (limits) as

$$a \pm t_{0.05}s_a \quad \text{and} \quad b \pm t_{0.05}s_b$$

and also draw the 95% confidence limits (bands) of the population regression line as

$$\hat{y}_i \pm t_{0.05}s_{\hat{y}}$$

which uses the sample standard error of a single sample prediction of $\hat{y}_i$.

Example 5

Thermocouples are common transducers used for measuring temperature. In order to use a thermocouple, the output voltage must be related to the temperature at which the voltage was obtained. Sample data for a thermocouple are given as:

mV (x)	T (y)
3.25	100.0
2.59	80.0
1.64	61.0
0.97	41.5
0.47	30.0
0.053	20.0
-0.75	0.0

Note that there must be measurement errors in both voltage and temperature, but for least squares analysis, we must assume that there is only error in the measurement of temperature. Some times, this assumption will be a good one, others not. For instance, if we measure an ice water mixture, we can be pretty sure that the temperature is in fact 0.0 °C. For calibration purpose, we want to use voltage as the independent variable. Hence in

[1] Explained on p. 164 in Snedecor and Cochran.

the least square fit, we only assume that temperature measurement follow a normal distribution. Calculating the sums required to get *a* and *b*:

$\Sigma(x^2)$	Σxy	Σx	Σy	$(\Sigma x)^2$
21.68	687.36	8.21	332.5	67.48

and applying the equations for *a* and *b* and with $n = 7$, we find

$$\Delta = (7)(21.68) - 67.48 = 84.25,$$

$$a = \frac{(21.68)(332.5) - (8.21)(687.36)}{84.25} = 18.5\ °C,$$

and

$$b = \frac{(7)(687.36) - (8.21)(332.5)}{84.25} = 24.7\ °C/mV$$

The values of *a* and *b* do not complete the picture. We also need to identify the error in our estimates of *a* and *b*. First we need to calculate a few more quantities from our data: the means of the voltages and temperatures, $\bar{y} = 47.5$ °C, $\bar{x} = 1.17$ mV, and the square of the deviations from the means:

$$\Sigma(x_i - \bar{x})^2 = 12.04 \text{ and } \Sigma(y_i - \bar{y})^2 = 7349.5.$$

Now, the sample variances:

$$s_{yi}^2 = \frac{7349.5 - (24.69)^2(12.04)}{7 - 2} = 2.47$$

$$s_b = \sqrt{\frac{(7)(2.47)}{84.25}} = 0.45$$

and

$$s_a = \sqrt{\frac{(2.47)(21.68)}{84.25}} = 0.80$$

With $t_{0.05} = 2.447$ for df = n – 1 = 6, the confidence intervals can be calculated. As an example, $\hat{y}$ is the value of y calculated using the a and b calculated above. The upper and lower confidence limits are calculated, for example, when x = 3.25 mV and $\hat{y} = 98.77$ °C.

$$s_{\hat{y}} = \sqrt{2.47\left(\frac{1}{7} + \frac{(3.25 - 1.17)^2}{12.036}\right)} = 0.896$$

and the confidence limits are $\hat{y} = 98.77 \pm (2.447)(0.896) = 96.0$ to 101.5 °C

Finally, the result is plotted in Figure E1.

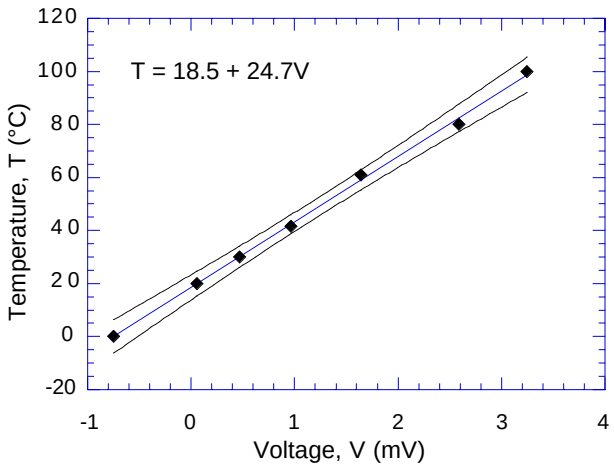


Figure E1. Thermocouple calibration with 95% confidence limits.

The misuse of the correlation coefficient

In the age of electronic calculators and computers, we have also come upon the era of boneheadedness with respect to the linear correlation coefficient. Yes, we can get this number with the simple push of a button, but what does it mean?

Let's see what this coefficient is all about. The *sample* correlation coefficient given by your almighty calculator is [1]

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{15}$$

or

$$r = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2]\,[n\Sigma y_i^2 - (\Sigma y_i)^2]}}$$

Assume that we have made measurements of pairs of quantities x_i and y_i and we want to know if the data correspond to a straight line of the form (we'll omit the ^ on the y)

$$y = a + bx$$

[1] If you come across names like the *Spearman's rank-difference coefficient* or the *Pearson product-moment coefficient*, have no fear. They are different ways to conceptualize (in mathematical terms of course) the degree of relationship between variables. If you have a chance to learn more about them, you may find that the Pearson coefficient is the slope of the regression of *y* on *x* when both are expressed in standard units (Z_x and Z_y). The Pearson *r* also shares similarities with eq. (15), which can be interpretation as the mean of the products of *x* and *y* in standard deviate units. These two statements may make no sense to you. I just hope that they may motivate you to read a real statistics text.

If indeed there exists a physical relationship between the two quantities *x* and *y*, we can equally well consider *x* as a function of *y* and ask if the data correspond to the following linear equation

$$x = a' + b'y$$

If there is no physical correlation between *x* and *y*, then either b or b' *must* be zero. If there is indeed a relation between these two variables, then there must also be a relation between the coefficients a, b in eq. (16a) and a', b' in eq. (16b). We can easily (as always!) show that

$$y = a + bx = -\frac{a'}{b'} + \frac{1}{b'}x$$

and equating coefficients, we obtain

$$a = -\frac{a'}{b'} \quad \text{and} \quad \mathbf{b = \frac{1}{b'}}$$

(The bold face is solely for visual effect.) The linear correlation coefficient, *r*, is defined as [1]

$$r = \sqrt{bb'} \tag{17}$$

Simple substitution will lead you to the messy formula given in eq. (15).

It should be obvious from (17) that $r = 0$ when there is no physical relation between *x* and *y*. Otherwise, the sign of *r* is simply a matter of whether the slope of the fit is positive or negative. When $r = 1$ or -1, many books use the term "perfect relationship." It is fine if we understand the perfect here means association.

◊ *What the correlation coefficient* ***does not*** *measure?*

When a correlation is perfect, there is no scatter. Of course, we can make accurate predictions of *y* with a given *x*, and we are tempted to use the coefficient as a measure of the "goodness of fit." However, strictly speaking, the coefficient of correlation is an index of relationship, *not* an indicator of the goodness of fit.

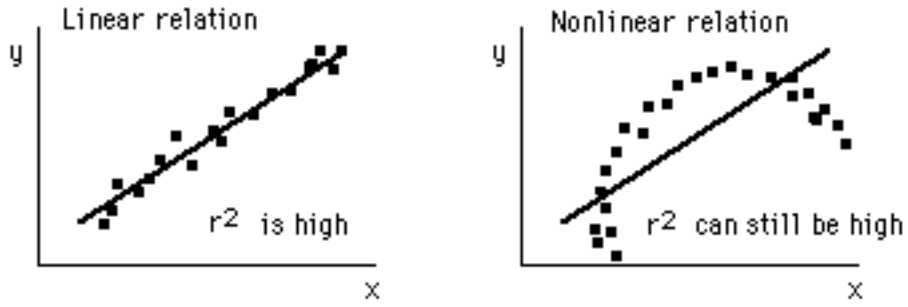
In fact, the correlation coefficient is not even a measure of the appropriateness of the straight line model. An (exaggerated) illustration is shown below.[2]

---

[1] Our derivation of *r* is not rigorous, but it is the easiest without having to use probability.

[2] Illustrations on how r can be misleading are in chapter 3 of R. Caulcutt, *Data Analysis in the Chemical Industry. Volume 1: Basic Techniques*, Chichester: Ellis Horwood, 1989, and chapter 6 of D.G. Kleinbaum and L.L. Kupper, *Applied Regression Analysis and Other Multivariable Methods*, North Scituate, Mass.: Duxbury Press, 1978.

In engineering or physical science, most of the experiments have a theoretical basis. So when we plot *y* versus *x*, we usually already have a preconceived idea (or bias in terms of statistics) that *y is* related to *x*. Otherwise, we wouldn't even be doing the experiment. And if we know that *y* should be related to *x* as a linear function, why do we have to bother with the correlation coefficient? Of course, the theory could be wrong, but it is extremely unlikely for undergraduate instructional experiments.



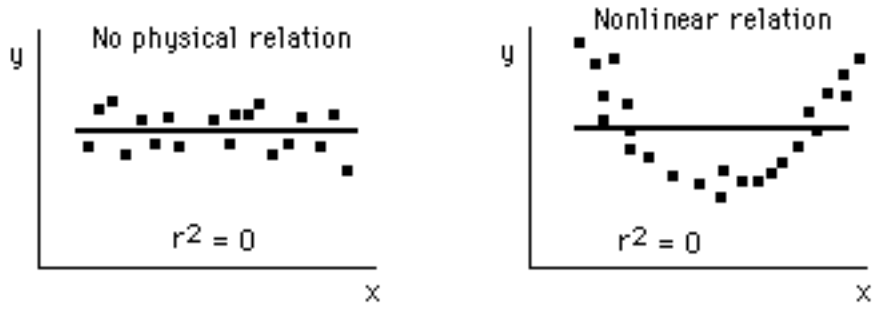
The value of *r* can also be distorted by highly scattered data points, which are called *outliners*. The use of the standard deviations of the slope, intercept, or the confidence limits is more appropriate for gauging the goodness of fit.

Furthermore, correlation is not causality. This is a point that we can easily miss because our regression analysis is almost always based on a physical model. Having a correlation is a necessary feature of a causal relation, but it is not sufficient to demonstrate causality. The pitfall of a causality lies in social science problems. For example, we may find students with a good vocabulary tend to have high overall SAT scores. Of course, this correlation does not mean having a good vocabulary alone is sufficient to achieve a high SAT score.

◊ *What does the correlation coefficient tell us?*

The correlation coefficient *r* is a **measure of association**. (*Do not* confuse this with the *goodness of fit.*) By association, we mean the existence of a physical relation between *x* and *y*. The lack of an association means that the value of one variable cannot reasonably be anticipated by knowing the value of the other variable.

The closer | *r* | is to unity, the stronger is the association between the two variables. The coefficient can be either positive or negative, depending on the slope, b or b', of the linear relation. If *r* is close to zero, there is little, if any, linear association between *x* and *y*, but a nonlinear correlation may exist.



It should be apparent that *r* in a way answers a yes/no question. We only get a statistically significant verdict that there is a linear relation with nonzero slope. How do we know that *r* is

close to zero? To test the hypothesis that the correlation coefficient is zero, we can use the *r* distribution.[1]

If you really want to use *r* to say something about your least-squares exercise, a more appealing interpretation is to use r^2 as a measure of "**percentage fit**,"

$$\text{percentage fit} = 100\ r^2$$

It can be shown that r^2 may be described *approximately* as the estimated proportion of the *variance of y* that can be attributed to its linear regression on *x*. The portion $(1 - r^2)$ is "free" from *x*.[2]

Let's say we have just least-squares fitted the number of drunk driving arrests (*y*) to the number of fraternity parties (*x*) with a correlation coefficient of 0.86. We could say that 74% (0.86^2) of the drunk driving is accounted for by the fraternity parties. The remaining 26% drunk driving arrests are due to other causes. These can be measurement errors or other variables not considered or controlled in the regression model.

Note that when we perform a least square calculation in engineering, we not only have a clear conviction that the two variables are related to one another, but we also can explain the relation with theories. Our experiments often are designed to verify theories. The question we address is the uncertainty involved if we use the regression model to make predictions. Our situation is very different from studies in social sciences and medical sciences. In these disciplines, there is seldom a physical model and the question at issue is indeed, "Are the two variables related?" [3]

Now that you have the equations for calculating the standard deviations for the slope and intercept, you will have no excuse for using the correlation coefficient to assess the goodness of fit in a simple two-variable regression analysis (*i.e.*, least-squares fit). Something you already should have learned from experience is that even though the data points are quite badly scattered, the correlation coefficient from your magic calculator still gives a value of *r* larger than 0.9. This also explains why we should *always plot the data points.* In a more rigorous analysis (available from most statistical packages such as IMSL, SPSS, SYSTAT, or RS1), you actually can provide the variance (or weighing factor) for each data point.[4] Consult a reference book before you use these canned packages.

---

[1] See chapter 7 in Bevington or chapter 10 in Snedecor and Cochran. We'll have to skip this. Notes are getting too long and you have the *t* test as the survival kit.

[2] The derivation is in chapter 10 of Snedecor and Cochran. You may find that r^2 is also called the **coefficient of determination**.

[3] On the other hand, **multiple correlation coefficients** are very useful in multiple variable regression analysis even in engineering. A correlation coefficient can be defined and evaluated for any given pair of variables and these coefficients can test for whether one particular variable should be included in the regression model to which the data are being fit.

[4] Bevington has the most consistent set of derivations in this respect.

Fitting a line through the origin

With many physical and chemical processes, we do expect, from theory, the regression line to go through the point (0,0). In this case, the regression model is as simple as

$$\hat{y} = \beta x + \varepsilon = bx \tag{18}$$

A couple of simple calculus steps will lead us to [1]

$$b = \frac{\sum x_i y_i}{\sum x_i^2} \tag{19}$$

and the sample (estimate) variance with now (n – 1) df

$$s_{y_i}^2 = \frac{\sum y_i^2 - \dfrac{(\sum x_i y_i)^2}{\sum x_i^2}}{n - 1} \tag{20}$$

and the sample variance of the slope *b* is

$$s_b^2 = \frac{s_{y_i}^2}{\sum x_i^2} \tag{21}$$

Now, there are two tests we'd like to perform. The first is to test the hypothesis that $\beta = 0$. The second is to ask whether the population regression line really goes through the origin. To answer this question, we have to repeat the two-parameter $y = a + bx$ regression to find the intercept, *a*, at $x = 0$. And we define the test statistic

$$t = \frac{a - 0}{s_a}, \quad \text{with} \quad df = n - 2$$

where s_a is calculated as before in the $y = a + bx$ regression analysis. The hypothesis is that H_O: $a = 0$. Our biased wish would most likely be that our measurements cannot reject the hypothesis (at a specified level of confidence).

ANOVA, Analysis of variance

We'll introduce a few more terms to help you read outputs from a statistical package. If *x* were useless in predicting *y*, our best guess of *y* ($\hat{y}$) would simply be $\bar{y}$ regardless of *x*. A table showing the contributions to total variability in the data and model estimation can be very useful, especially in multiple regression analysis.

In a simple linear regression, the total variation among the data y_i as measured by the sum of square $\sum(y_i - \bar{y})^2$ can be split into two parts:[2]

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 \tag{22}$$

---

[1] Well, you will have to do it if you don't believe it. It is just a simple homework problem that takes a couple of lines.

[2] The derivation, actually quite easy, is in Chapter 9, Snedecor and Cochran. For a more comprehensive introduction, see Chapter 12 in Rice.

where the first term on the right, $\Sigma(\hat{y}_i - \bar{y})^2$, is a measure of the total variation among the regression estimates and $\Sigma(y_i - \hat{y}_i)^2$ is the residual variation not accounted for by the regression.

The mean square of the total variation with df = n – 1, $\frac{\Sigma(y_i - \bar{y})^2}{(n-1)}$, is simply the sample variance of the data y_i.

The residual sum of squares $\Sigma(y_i - \hat{y}_i)^2$ is also called the chi-square (χ^2). The mean square of the residual sum of squares with df = n – 2, $\frac{\Sigma(y_i - \hat{y}_i)^2}{(n-2)}$, is simply the sample estimate variance, $s^2_{y_i}$ defined in eq. (11). The residual mean square (or estimate variance) is obviously a measure of how poorly or how well the regression line fits the data points.

The df of the regression sum of square is only 1 and the mean square of the regression remains $\Sigma(\hat{y}_i - \bar{y})^2$. If we take the ratio of the regression mean square to the residual mean square (also called the F test), we can show (from reading a text!) that

$$F = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{s^2_{y_i}} = \frac{b^2\Sigma(x_i - \bar{x})^2}{s^2_{y_i}} = \frac{b^2}{s^2_b} = t^2 \quad \text{for the hypothesis Ho: ß = 0} \tag{23}$$

You will find that most canned packages report these values under the acronym ANOVA. And since we flaunted the terms χ^2-square and F test and they are not used in their more proper definitions, we have one last business to do.

The χ^2 and F distributions

We finish this review with two definitions. We'll not go into the mathematical or computational details. Rather, the focus is on the meaning and interpretation which often are lost among the theoretical details in a text. So we really are explaining the χ^2 and F statistics. Their distributions can be found in statistical packages or in appendixes of textbooks, much like the *t* distribution.

The χ^2 distribution

The χ^2 (chi-square) statistic is a measure of the total difference between measurements and their corresponding expected values. The expected values can be taken from a theory, a model, or a hypothesis, and a common hypothesis is that the samples are randomly taken from a normally distributed population.

In mathematical terms, we define the chi-square statistic as

$$\chi^2 = \sum \frac{(f_i^{obs} - f_i^{ex})^2}{f_i^{ex}} \tag{24}$$

where f_i^{obs} is the observed or measured frequency of the *i*-th sample, and f_i^{ex} is its corresponding expected frequency. The square of the difference is divided by the expected frequency, and the chi-square is the sum of these squared differences for all the samples.

In the case of linear regression, the chi-square based on the residual sum of squares is more properly stated as

$$\chi^2 = \sum \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$$

where the expected value of each measurement is taken from the statistical regression model.

The F distribution

The F test uses a ratio between two variance estimates.[1] The F ratio can be defined as

$$F = \frac{s_b^2}{s_w^2} \tag{25}$$

where the numerator and denominator are the population variances as *estimated* from two different measures of observed variability, and they are denoted by the *b* and *w* subscripts.

One scenario where the F tests is applicable is as follows: Say we have *n* tests (or test groups) that are supposed to yield the same result. Of course, there are variabilities in the measurements and each test yields slightly different means. We now want to know if there is any significant difference *anywhere* among these tests. (Another option is to perform paired *t* tests for each chosen pair, but we would soon run out of patience if we do that.) Here, we compute the F ratio with the numerator being the variance estimated from the means of the *n* groups (hence the *b* for between groups), and the denominator is the variance estimated from the variability of individual scores *within* their groups (hence the *w*).

So what does this ratio tell us? The variance of scores within a test group reflects how individuals may deviate from within their own group mean. Each test group is different and the numerator is the variance of the means of the different groups. If all the test objects are the same, the variability should be the same whether it is between or within groups, and the F ratio should be 1. If indeed there are differences among the test groups, the variance among the groups will be larger and $F > 1$.

From the perspective of analysis of variance, and in the words of John Phillips, “every individual deviation from the ‘grand’ mean (of all sampling in the *n* test groups) is made of two components: (1) the deviation of an individual score from its group mean, and (2) the deviation of that group mean from the grand mean.” In linear regression, the F test is based on the two components as defined in eq. (22) and the ratio of the regression mean square to the residual mean square as shown in eq. (23).

Do the χ^2 and F ration make much sense to you yet? Probably not. Do not panic. These concepts are never easy to digest. To fully understand their usage, we need some extensive numerical examples, which will be left to your statistics courses. These review notes are long enough.

---

[1] The F test is named after Sir Ronald Fisher, who developed the concept of the analysis of variance.