8 Hypothesis Testing

8.1 Introduction

The Ionian GBCE™ confidence limits for the standard deviation in §7.24 came very close to asking whether the sample standard deviation could have come from a population whose parametric standard deviation was greater than 2. Hypothesis testing in statistics is usually in the form of the question, "Could the results we observe in our sample have occurred by chance variation in sampling alone if one or more parameters had specific values?" The hypothesis that defines the baseline against which the sample is evaluated is called the *null hypothesis* (H0 or H_0 in most textbooks) and the particular hypothesis that will be accepted if we reject the null hypothesis is called the alternative hypothesis (most commonly H1, H_1 or H_a).[88]

Typically, the question of *interest* will be represented by the alternative hypothesis. As illustrated in the following examples, note how consistently what is interesting to the analyst is the alternative hypothesis in the following examples of some questions we might encounter and the corresponding statistical hypotheses that might be framed:

- An accountant doing an audit is becoming suspicious of the figures shown in the books of a big company called Unron; she extracts the data from several hundred transactions and wants to know if the frequencies of the ten digits (0, 1,…9) in the last portions of the entries are equal (radical deviation from equality would suggest that the numbers were fraudulently invented, since people aren't very good at making up numbers that fit the uniform probability distribution).
    - H0: the frequencies *are* all 0.1;
    - H1: the frequencies *are not* all 0.1.
- A stock broker has become interested in the performance of the shares for Macrohard Corporation; he wants to know if the data for the last three years support the view that the growth rate is at least 6% per year. If it is, he will recommend to a client interested in long-term investments that the investment fits the client's profile.
    - H0: the regression coefficient is *less than* 6% per year.
    - H1: the regression coefficient of stock price versus year is *6% per year or more*;
- A network management team needs to know if a new heuristic compression algorithm is working *better* than the old Lempel-Zev-Welch (LZW) algorithms that have been in use for decades. They measure the compressed file sizes for a wide range of files using both types of algorithms and compare the results.
    - H0: there is *no difference* between the compression ratios of the LZW compression methods and the new heuristic algorithms OR the heuristic algorithms are *not as good as* the LZW methods.
    - H1: the heuristic algorithms are *better* than the LZW algorithms.

[88] This text uses H0 and H1 to avoid constantly having to typeset subscripts, especially in Excel, where subscripts are a real pain to create. In exercises, students are exposed to both H1 and Ha for practice.

- A manufacturer of trans-temporal frammigers is investigating the effects of besnofring modulation on the accuracy of the time-tuning mechanism. The investigators apply seven different levels of besnofring modulation to the frammigers while transporting samples to ten different time locations each. They analyze the 70 measurements to see if there are any effects of the besnofring modulation.
    - H0: there are *no effects* of differences in besnofring modulation level on the accuracy of time-tuning.
    - H1: there *are* differences in accuracy of time-tuning among the groups exposed to different levels of besnofring modulation.
- A marketing firm has three different variations on an advertising campaign. They are trying them out in six different regions of the target market by counting the number of answers to questions using a Likert scale (1: strongly disagree, 2: disagree… 5: strongly agree).
    - Are there differences among the ad versions in the way people respond in the six regions overall?
        - H0: there are *no* main (overall) effects of ad versions on responses;
        - H1: there *are* main effects of ad versions on responses.
    - Are there differences among the regions of the market in the way people respond?
        - H0: there are *no* main effects of region on responses;
        - H1: there *are* main effects of region on responses.
    - Are there any regional variations in the way people respond to the different campaigns
        - H0: there are *no interactions* between the ad variations and the regions of the market in the way people respond;
        - H1: there *are* interactions between the ad variations and the regions of the market.
- An investor is looking at two different manufacturers of trans-temporal frammigers as potential investments. One of the steps in due diligence is to examine the reliability of quality control of the two factories' production lines by comparing the variances of the products.
    - H0: there is no difference in the variances of the two production lines;
    - H1: there is a difference in the variances of the two production lines.
- A system manager needs to know if a particular department's growth rate in disk space utilization is really faster than all the other departments' growth rates.
    - H0: the regression coefficient of disk space over time for the suspect department is *not different* from the other departments' regression coefficients or it is *slower*.
    - H1: the regression coefficient of disk space over time for the suspect department is *greater* than the other departments' regression coefficients.

In general, the null hypothesis, H0, is the one that posits no effect, no difference, no relationship, or a default state. H1, the alternative hypothesis, is the one that posits an effect, a difference, a relationship, or deviation from a ho-hum uninteresting state. There is nothing absolute or inevitable about the framing of H0 and H1: the decisions depend on the interests of the investigator.

8.2 Are the Variances of these Two Samples the Same?

Although most statistics textbooks start with comparisons of the means, it will be very useful for you to know about testing variances – the test is used in analysis of variance (ANOVA), which is a central technique in applied statistics. Here we begin the study of hypothesis testing by looking at variances.

Statisticians have determined that when we repeatedly take two samples from the same population and compare their variances, the ratio, called an F-statistic, follows a distribution called the F-distribution with

$\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom. The degrees-of-freedom parameters apply to the numerator and the denominator of the ratio.

For the ratio of two sample variances, s^2_1 and s^2_2 computed the usual way from samples of size n_1 and n_2, we compute

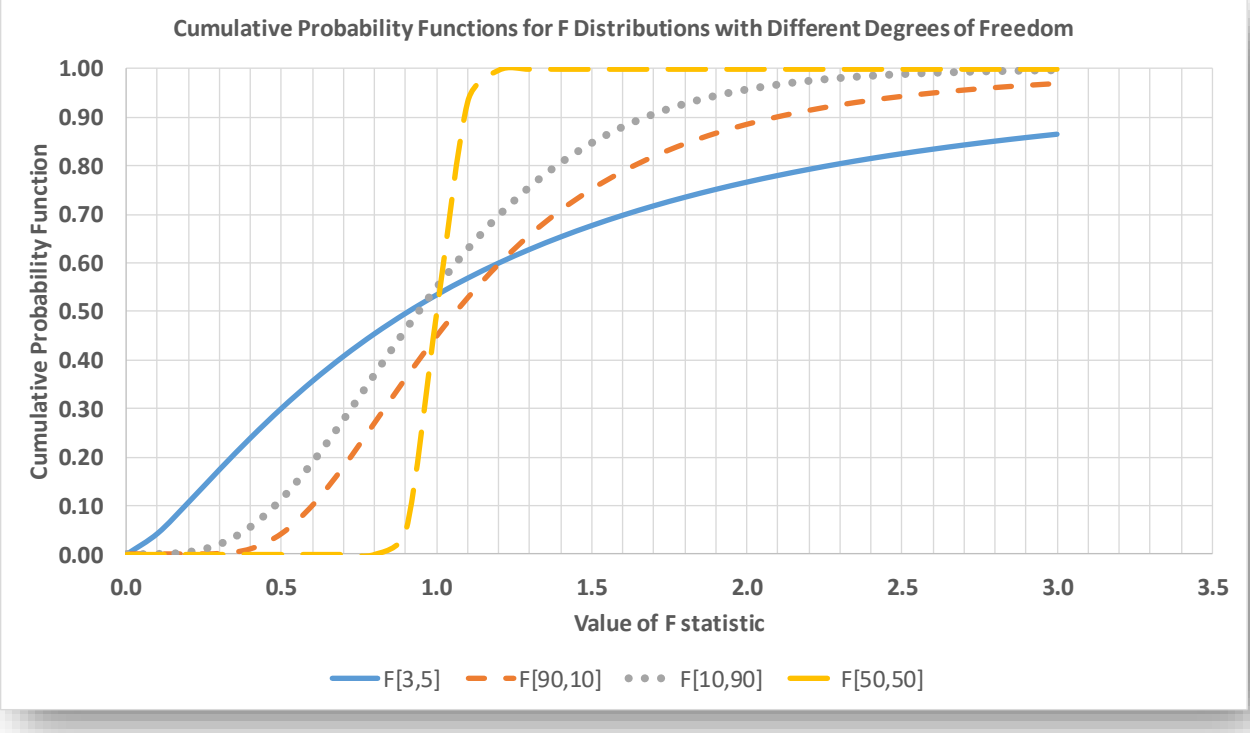
$$F_{\nu_1 \nu_2} = \frac{s^2_1}{s^2_2}$$

This *F-ratio* follows the F distribution with ν_1 and ν_2 degrees of freedom.

Each combination of degrees of freedom has a different shape, as shown in Figure 8-1, which graphs cumulative probability functions for four combinations of n_1 and n_2 degrees of freedom.

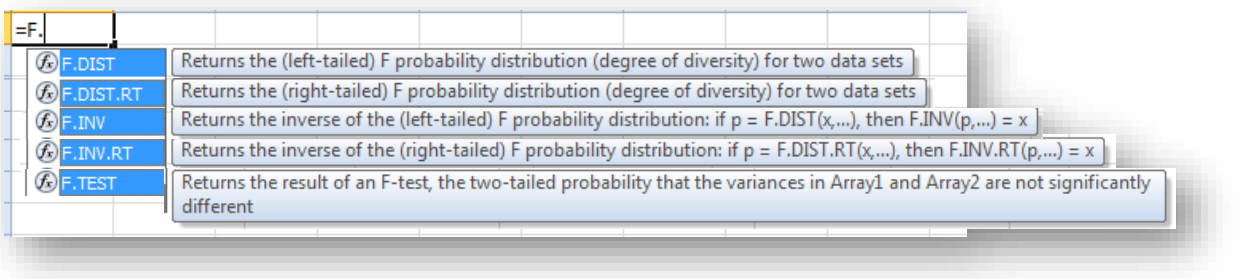
The F-test for the equality of two sample variances (or of any other quantity that follows the same rules as variances) consists of computing the ratio the sample variances and determining how often such a large value of F or larger would occur by chance alone (i.e., through random sampling) if both samples actually came from the same population or from two populations with identical parametric variances.

Figure 8-1. Cumulative probability functions for various F distributions.



By convention, we create the F-ratio by dividing the larger sample variance by the smaller sample variance so that F will be 1 or larger, allowing us to use F.DIST.RT (=FDIST in earlier versions) to compute the right-hand tail of the F- distribution for the appropriate pair of degrees of freedom. The EXCEL function =F.DIST.RT(F, n1, n2) provides the probability that the value F (and any values larger than the one observed) of the F-ratio would occur by chance if the samples came from the same population or if they came from two populations with identical parametric variances. The F-distribution functions are shown in the composite image in Figure 8-2.

Figure 8-2. Excel 2010 F-distribution functions.



For example, =F.DIST.RT(2.5,10,100) = 0.010; thus the probability of obtaining an F-ratio of 2.5 or larger by chance alone when the degrees of freedom of the numerator and denominator are 10 and 100, respectively, is 0.010.

If for some reason you are given an F-ratio that is smaller than 1, you can use the **=F.DIST** function to calculate the left-tail probability. For example, if the F-ratio were inverted, giving **=F.DIST(0.4,100,10,1)**,[89] the result would be 0.010, as it should be.

There are three possible sets of hypotheses that you might want to test with an F-test:

- H0: $\sigma^2_1 = \sigma^2_2$ and H1: $\sigma^2_1 \neq \sigma^2_2$
- H0: $\sigma^2_1 \leq \sigma^2_2$ and H1: $\sigma 21 > \sigma^2_2$
- H0: $\sigma^2_1 \geq \sigma^2_2$ and H1: $\sigma 21 < \sigma^2_2$

Suppose we are interested in the political commitment of two different groups of voters in the Martian general elections, the Mars First Coalition secessionists and the Earth Forever Patriots irredentists. One measure of their commitment is the variability of their responses to ideological statements reflecting each movement's core values. A University of Schiaparelli political science group develops test instruments and applies them to 1,228 Mars Firsters (secessionists) and to 1,175 Earth Forever (irridentists) members. Their results show that the variance of the overall scores are 38.2 for the secessionists and 35.0 for the irredentists. Are there grounds for believing that the two groups have a difference between their parametric variances?

Calculate the F-ratio that is greater than 1:

$$F_{[1227,1174]} = 38.2/35.0 = 1.091429$$

Then

$$P\{ F_{[1227,1174]} \geq 1.091429 \mid \sigma^2_1 = \sigma^2_2 \} = \text{F.DIST.RT(1.091429, 1227, 1174)} = 0.06504$$

We can read the line above as "The probability of obtaining an F-ratio of 1.091429 or larger by chance alone using these sample sizes if the two population variances were equal is 0.0650."[90]

[89] For unknown reasons, Microsoft decided to include the *cumulative* parameter (1=cumulative, 0=density) in the =F.DIST function but not in the the =F.DIST.RT function.

[90] P{ a | b} is called a *conditional probability* and is interpreted as "The probability of observing *a* given that *b* is true."

8.3 Levels of Statistical Significance and Type I Error: Rejecting the Null Hypothesis When it is Actually True

At this point, a reasonable question is "So what if the probability of the F-ratio is 0.0650? What does that mean to us? Are the parametric variances different or aren't they?"

A conventional answer points to four agreed-upon levels of statistical significance in hypothesis testing. We conventionally refer to the probability of obtaining the observed results by chance alone if the null hypothesis is true as $P\{H0\}$ and even more often simply as p. We also call p the probability of Type I Error.

Type I Error is *rejecting the null hypothesis when it is actually true.*

We agree by convention that

- If $p > 0.05$
  - We accept H0,
  - Reject H1, and
  - Say that the results are *not statistically significant* and
  - Follow the P{H0} (or p) with the letters *ns* to show non-significance; e.g., "The probability of encountering such a large statistic or larger if the parametric variances were the same is 0.0650, which is not statistically significant at the 0.05 level of significance. We therefore accept the null hypothesis of equal variances for these two samples.
- If $0.01 < p \leq 0.05$
  - We reject H0,
  - Accept H1, and
  - Say that the results are statistically significant or statistically significant at the 0.05 level.
  - The P{H0} (or p) is typically marked with one asterisk (*) to show this level of significance; e.g., P{H0} = 0.0472*
  - The most explicit formulation of the results is something like this: "The test statistic of 34.98 has a probability of 0.0472* of occuring by chance alone even if the samples had the same parametric value of this statistic. The results are statistically significant at the 0.05 level of significance and we reject the null hypothesis of equality of the parametric statistics."
  - In practice, we would be more likely to write briefly, "The parametric statistics for these samples are significantly different at the 0.05 level of significance (p = 0.0472*)."
- If $0.001 < p \leq 0.01$
  - we reject H0,
  - accept H1, and
  - Say that the results are statistically highly significant or statistically significant at the 0.01 level.
  - The P{H0} is typically marked with two asterisks (**) to show this level of significance; e.g., P{H0} = 0.00472**.
  - The verbose formulation could be, ""The test statistic of 87.02 has a probability of 0.00472** of occuring by chance alone even if the samples had the same parametric value of this statistic. The results are statistically significant at the 0.01 level of significance and we reject the null hypothesis of equality of the parametric statistics."
  - In practice, we would probably write something like, "The parametric statistics for these samples are highly significantly different at the 0.01 level of significance (p = 0.00472**)."

- If $p \leq 0.001$
  - We reject H0,
  - Accept H1, and
  - Say that the results are statistically extremely significant or statistically significant at the 0.001 level.
  - The P{H0} is typically marked with three asterisks (***) to show this level of significance; e.g., P{H0} = 0.000472***.
  - The verbose formulation could be, "The test statistic of 109.53 has a probability of 0.000472** of occuring by chance alone even if the samples had the same parametric value of this statistic. The results are statistically significant at the 0.001 level of significance and we reject the null hypothesis of equality of the parametric statistics."
  - In practice, we would probably write something like, "The parametric statistics for these samples are extremely significantly different at the 0.001 level of significance (p = 0.000472***)."

So going back to the study of Martian colony politics, are the two political movements similar or different in their variability on their ideological position?

The way a statistician would normally respond to the test results would be to say that the results of the study were not statistically significant, since the probability of rejecting the equality of variances even though they might be the same was about 6%, above the normal cutoff point of 5% for statistical significance. However, taking into account the size of the samples (respectable) and the closeness of the *p*-value to the limit for significance, a statistician would also add that the results are *highly suggestive* even though not statistically significant and that the problem could bear further study.

The issue of whether experiments or studies are repeated if the initial results don't fit preconceptions or preferences is a, you should pardon the term, significant problem in statistics. Even without a theoretical analysis, it must be clear that repeating studies until one gets the result that's desired – and ignoring the contrary results of the earlier studies – is surely going to bias the results towards the preconceived goal. A reasonable approach must repeat experiments if they are close to a minimum regardless of whether they are on one side or another. Thus a good experimental protocol might include "The experiment will be repeated if the p-value is from **0.02** through 0.08." What *won't* work is "The experiment will be repeated if the p-value is from **0.05** through 0.08."

Here's another example of a test of variances, this time drawn from computer science.

A musician working with machine intelligence has created a computer program that analyzes samples of music from specific composers – sometimes thousands of compositions – and creates complex algorithms reflecting the composers' patterns, including the degree of spontaneity and originality, in their music. Using these algorithms, the program interacts with its operator to construct new music that can be in the style of the original composer or that can be wholly original. The musician is studying listeners' response to the music and is interested in knowing if the range of responses is different when they listen to the original composers' music compared with the range of responses when they listen to the synthesized music. Thus

$$H0: \sigma^2_1 = \sigma^2_2 \quad \text{and} \quad H1: \sigma^2_1 \neq \sigma^2_2$$

The researcher collects data on 40 people's responses and finds that the F-ratio of the variance of the responses to the original music compared to the variance of the responses to the synthesized music is 1.83 with 39 and 39 degrees of freedom and $P\{H0\} = 0.0314*$. We can say "There is a statistically significant greater variance of the responses to the original music compared with the variance of the responses to the synthesized music ($p = 0.0314*$)".

As we have seen, we often encounter assertions of the form, "If the null hypothesis is true, the chance of observing this statistical result or one even more extreme is 0.042*, which is statistically significant at the 0.05 level of significance."

But what if we are unlucky? What is the probability of observing the results we obtained if the null hypothesis is actually true but we have stumbled into the improbable case? We simply assert that if the probability that the null hypothesis is ***p***, then the risk of *rejecting* the null hypothesis when it is *true* is simply ***p***. So in the example above, the probability ***p*** of encountering the observed results if H0 is true is 0.042 and the probability that we will be wrong in accepting the null hypothesis is exactly that: 0.042.

8.4 Type II Error: Accepting the Null Hypothesis when it is Actually False.

In general, if we choose the α level of signficance, we have to understand that we will be *wrong* on average in α of the statistical decisions we make that accept the null hypothesis. Thus if we choose α = 0.01, there is a 1% chance that we will encounter the deviant results that will trick us into rejecting the null hypothesis when it is true.

It follows that if we increase the severity of our test criterion – e.g., if we use α = 0.001 as our limit for rejecting the null hypothesis, there is an increasing risk that we will *accept* the null hypothesis when it is actually false. We call this kind of error *Type II Error*.

Figure 8-3. Type I and Type II Errors.

Decision Under Uncertainty	H0 TRUE	H0 FALSE
Accept H0	CORRECT	TYPE II ERROR
Reject H0	TYPE I ERROR	CORRECT

The calculation of the probability of Type II Error (sometimes symbolized β) is complicated because it's impossible to calculate theoretical probabilities of the observed values without specifying an explicit value for the null hypothesis. If H0 assumes equality of means, there is an infinite number of versions of H0 in which the means differ.

Statisticians have developed methods for estimating the probability of Type II Error, but the topic is suitable for a more advanced course than this introduction.

It is enough for the time being for students to be aware that there is always a possibility of being wrong in our *decisions under uncertainty*.

8.5 Testing a Sample Variance Against a Parametric Value

It is possible that you may have to test the variance of a specific sample, s^2, against a known or assumed parametric value, σ^2. You may recall the example in §7.24, where we developed a confidence interval for the parametric variance based on a sample statistic: the situation involved production of 100 gram bottles of a beetle-carapace extract at an Erewham plant, where the rules were to start investigating the production line if the standard deviation of the sample exceeded a parametric limit of 2 grams. One approach was to compute confidence limits; now you can see that another approach is to test the hypothesis that the parametric standard deviation of the population from which a a sample has been extracted is equal to 2 grams or less. If so, no problem; if not, problem! Exceeding the maximum allowable variance could lead to throwing the entire production into the compost bin.

We have to convert our standard deviations into variances to be able to calculate the test value. So $s^2 = 0.8^2 = 0.64$ and $\sigma^2 = 2^2 = 4$ and the null and alternative hypotheses in this case would be

$$H0: \sigma^2 \leq 4 \qquad \text{and} \qquad H1: \sigma^2 > 4.$$

In §7.24, you were told that the quantity

$$\mathbf{x = (n - 1)s^2/\sigma^2}$$

follows a chi-square distribution with df = (n – 1) degrees of freedom. This quantity thus becomes the basis of a hypothesis test for sample variances compared to parametric variances.

We compute the quantity x and compare it to the $\chi^2_{[n-1]}$ distribution to determine the likelihood that the x value (or a value even more unexpected – i.e., larger) could have occurred by random sampling alone if the sample really came from a population with the defined parametric variance $\sigma^2 = 4$.

Let's continue with our beetle carapace example and look at the results of the sample discussed in §7.24,. The quality-assurance test took 25 bottles of the hugely popular *Grandiloquent Beetle Carapace Extract* (GBCE™) from Vega II on a particular day in 2281and found the sample standard deviation to be 0.8 gm. Summarizing the situation, we have

$n = 25$ $s = 0.8$ and so $s^2 = 0.64$ $\sigma = 2$ and so $\sigma^2 = 4$

Therefore

$$x = (n - 1)s^2 / \sigma^2 = 24*0.64/4 = 3.840$$

and this quantity is distributed as the $\chi^2_{[24]}$ distribution if the data are the result of random sampling from a population (production line) with a parametric variance of 4 (parametric standard deviation of 2).

Using the CHISQ.DIST.RT function in EXCEL 2010,

P{H0} = CHISQ.DIST.RT(3.94,24) =0.9999988302 ≈ 1

in other words, it is *almost certain* that this sample could have been drawn from a population (the production for that day) with a parametric variance of 4 (i.e., a parametric standard deviation of 2). Therefore, we accept the null hypothesis: $\sigma^2 \leq 4$. Production of beetle carapace extract can continue without risk of conversion of the product to compost (what a relief)!

Just for practice, what if the sample standard deviation s had actually been 3.0 instead of 0.8? Then $s^2 = 9$ and

$$x = (n - 1)s^2 / \sigma^2 = 24*9/4 = 54 \text{ and}$$

P{H0} = CHISQ.DIST.RT(54, 24) = 0.000426***

Technically, we would write that "The probability that we could obtain a sample with standard deviation of 3 or more if the parametric standard deviation were 2 or less is only 0.0004*** which is extremely statistically significant at the 0.001 level of significance."

8.6 Are the Means of These Two Populations the Same?

In every field of investigation, people need to know if the results of their measurements in two different samples could have come from the same population . Are these samples from the same population and therefore different due to random sampling alone? Or are they from two different populations whose parametric means differ? One of the most powerful tools in applied statistics is the analysis of variance, or ANOVA, which builds on what we have learned about testing the equality of sample variances. Another widely-used tool is the t-test of the equality of two statistics that follow a Normal distribution.

8.7 ANOVA for Comparing Means of Two Samples

Imagine that we have two samples of size 30 that we know are actually drawn from the *same population of measurements* of a production line of 10" radius tires for very small environmentally-clean cars powered by pet rabbits. We need to find out if the means of the two samples reflect differences in the parametric means of the populations sampled or if we can accept that the means as equal (H0).

The mean radius and variance for each sample are as follows (the unusual precision is provided so readers can do their own computations):

Figure 8-4. Mean and variance of two samples.

Samples	Mean Radius	Variance of radius
Sample 1	9.992921717	0.08067098
Sample 2	10.03033012	0.08771458

Before we can use ANOVA to test the significance of the observed difference between the means, we should satisfy ourselves that one of the key assumptions of ANOVA is satisfied: the parametric variances of the samples must not differ. Using our usual notation, we must test

H0: $\sigma^2_1 = \sigma^2_2$ and H1: $\sigma^2_1 \neq \sigma^2_2$

We know, as godlike beings, that there are really no differences in the parametric statistics of these two samples. They are just samples from the same population. The statisticians working with the data, however, don't know this: they have to manage with probability calculations.

Noticing that the variance for the second sample is bigger than that of the first sample, we compute the F-test for equality of the parametric variances of the populations from which these samples are drawn as

$F_{[29,29]} = s^2_2/s^2_1$ to generate a value > 1 for the test.

$F_{[29,29]} = 0.08771458/0.08067098 = 1.087$

How likely is it that we could observe an F-ratio that large or larger if it were true that H0: $\sigma^2_1 = \sigma^2_2$?

We compute

= F.DIST.RT(1.087, 29, 29) = 0.412ns

and thus conclude that there is no reason to reject H0: $\sigma^2_1 = \sigma^2_2$. We may proceed with the test of the means.

Now suppose we combine the two samples into one big sample with n = 60 and compute the mean and the variance of the pooled sample where the means are the same? Without going into details, the results would be as follows:

Mean radius (pooled): 10.01162592 and **Variance of radius (pooled): 0.083121563**

There is nothing surprising about this: after all, from our godlike perspective as creators of the example, we know perfectly well that the samples really are from the same population. Pooling the samples would be the same as simply taking a random sample of size 60 instead of two samples of size 30. The mean of the pooled data isn't much different from the two-sample means and the variance is pretty much like the separate variances.

But what if we alter the thought-experiment and imagine that the second sample actually comes from a different production line: one making 17" radius tires (7" bigger than the tiny 10" radius tires) for gas-guzzling monsters with macho supercharged engines and, ah, virile mag wheels? What would happen to the statistics we just computed?

Adding 7" to the existing Sample 2 to create a synthetic Sample 3 for our thought experiment produces the following results:

Figure 8-5. Sample statistics for tires including new Sample 3 modified by adding 7 to every value in Sample 2.

Samples	Mean Radius	Variance of radius
Sample 1	9.992921717	0.08067098
Sample 2	10.03033012	0.08771458
Sample 3	17.03033012	0.08771458

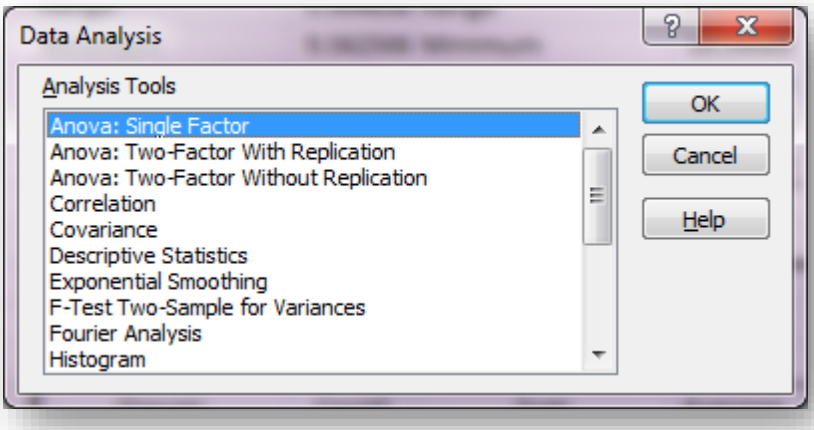
The mean of Sample 3 in Figure 8-5 is exactly 7" larger because we deliberately added 7" to every one of the observations we already had for Sample 2; it's no surprise that the mean increased by exactly 7". What may surprise you is that the variance of Sample 3 is exactly the same as the variance of Sample 2, but if you think about it, variance is computed as a function of the average deviations of the values in a set from their mean, so there was no increase in variability of the data after the artificial shift in the mean. We just shifted all the Sample 2 data up by 7"; we didn't increase the spread of the data at all in the new Sample 3.

Without doing a computation, we can say intuitively that *the average variance within the groups is the same regardless of the difference in the means of the two groups*. We won't delve too deeply just now into how to compute the average variance within the groups – that's coming in a few paragraphs.

But for now we compute the *overall variance of the pooled data* including Sample 1 and Sample 3: the mean is 13.51162592 and the variance is 12.67389725. The variance of the pooled data when the three samples have different means is bigger (~13.5) than when the two samples have the same mean (~0.08).

It turns out that the one-way analysis of variance (ANOVA) is based on something very similar to this kind of reasoning. After activating the Data | Data Analysis tools menu, we can use the function Anova: Single Factor (Figure 8-6).

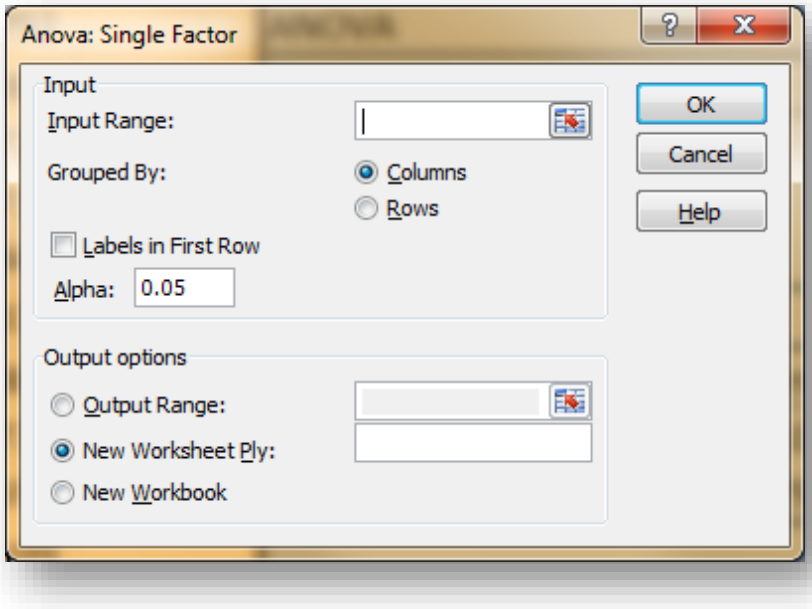
Figure 8-6. Choosing the ANOVA for testing differences of means classified by a single factor.



The concept of *single factor* refers to the classification of groups according to only one criterion; e.g., sample number or team or company or whatever we are interested in defining as our groups of interest.

Figure 8-7 shows the menu for entering the locations of the input data and the output in EXCEL 2010.

Figure 8-7. Menu for ANOVA single factor in Excel 2010.



Later, you'll study two-factor ANOVAs that classify groups by two criteria (factors) such as gender and geographical location, or company and size, or weight and height, or team and year….

Figure 8-8 shows the output for our example, starting with Samples 1 and 2 that were similar in their means:

Figure 8-8. ANOVA for tire data – samples 1 & 2 (means not different).

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Sample 1	30	299.7876515	9.992922	0.080671
Sample 2	30	300.9099037	10.03033	0.087715

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	0.020991	1	0.020991	0.249319	0.619444
Within Groups	4.883181	58	0.084193		
Total	4.904172	59			

Looking at Figure 8-8, you can see that the F-test in this case is the ratio of two variance-like measures of variability (called the *Mean Squares* and abbreviated *MS*).

- The Between Groups MS
  - Is a function of the differences in the means of the two (or more in general) samples,
  - Is in the row labeled *Between Groups* in the column mark.ed *Source of Variation*)
  - Is often printed as MS_{groups} in discussions of the results.
- The Within Groups MS
  - Is based on the average variations inside both samples
  - Is often written MS_{within}
  - Is often called MS_{error}.

The null and alternate hypotheses here are expressible in three ways that imply the same thing:

- H0: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$
- H0: $MS_{groups} = MS_{within}$ H1: $MS_{groups} \neq MS_{within}$
- H0: $F \leq 1$ H0: $F > 1$

As you can see in , the F-ratio of 0.249319[91] is not statistically significant; there is no reason to reject the null hypothesis that the means of samples 1 and 2 are the same. The p-value of ~0.6 means that the chances of getting an F-ratio that large or larger with 1 and 58 degrees of freedom if the null hypothesis were true is about 3 out of 5 sampling experiments.

To check the calculation of F (just to help make sense of it), one can calculate =F.DIST.RT(0.249319,1,58) which sure enough gives 0.61944 just as the ANOVA table in Figure 8-8 shows.

[91] Usually three significant figures are enough when publishing an F-test ratio, so we would normally adjust the number of decimal places to show *0.249*.

Now let's look at the ANOVA for Samples 1 and 3:

Figure 8-9. ANOVA for tire data – samples 1 & 3 (different means created in demo data set).

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Sample 1	30	299.7876515	9.992922	0.080671
Sample 3	30	510.9099037	17.03033	0.087715

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*
Between Groups	742.8768	1	742.8768	8823.521	4.5E-65
Within Groups	4.883181	58	0.084193		
Total	747.7599	59			

Because we deliberately changed the mean of Sample 3 in this illustrative example,[92] the MS_{groups} is huge: >742! However, because of the artificial nature of the modification, the MS_{within} is *exactly the same* as in Figure 8-8 comparing samples 1 and 2. The F-ratio (MS_{groups} / MS_{within}) is now enormous (>8823).

There is no reasonable doubt that we can *reject* the null hypothesis of equality of the means of samples 1 and 3 in this contrived example: the chances of getting such as large F-ratio (>8823) with 1 and 58 degrees of freedom are negligibly small (4.5E-65). The result is thus described as *extremely statistically significant*.

So let's recapitulate what you've seen in this section.

- ANOVA is based on the concept that if everything is the same in samples, a form of variance for individual samples versus the variance of the entire set of data should be similar.
- ANOVA depends strongly on the assumption that all samples being compared *have the same fundamental (parametric) variance*. We call this the assumption of *homoscedasticity*.[93]

[92] We added 7 to every value.

[93] Sokal & Rohlf write, "Synonyms for this condition are *homogeneity of variances* or ***homoscedasticity,*** a jawbreaker that makes students in any biometry class sit up and take notice. The term is coined from Greek roots meaning 'equal scatter.'" (Sokal and Rohlf, Biometry: The Principles and Practice of Statistics in Biological Research 2012) p. 414.

8.8 The Model for Single-Factor ANOVA

The underlying model for this ANOVA can be written as follows:

$$Y_{ij} = \overline{\overline{Y}} + \alpha_i + \varepsilon_{ij}$$

where Y_{ij} is the jth observation in group i of the comparison;

$\overline{\overline{Y}}$ is the overall (global) mean of all the observations;

α_i is the average main effect of being in group i (i.e., $\overline{Y}_i - \overline{\overline{Y}}$);

ε_{ij} is the residual variation for the jth observation in group i (i.e., $Y_{ij} - \overline{Y}_i$).

SS in the ANOVA table stands for *Sum of Squares* and, as you have seen already, *df* stands for *degrees of freedom*. A *Mean Square* (*MS*) is *SS*/*df*. The MS_G is the *Mean Square Among (or Between) the Groups* ; for *a* groups, MS_G has $df_G = a - 1$ degrees of freedom. $MS_G = SS_G / df_G$. Similarly, MS_W is the *Mean Square Within Groups* (also known as the *Mean Square Error*, the *Residual Mean Square Error*, and the *Unexplained Mean Square Error* abbreviated MS_E) and has $df_W = \Sigma n - a$ where Σn is the sum of all the sample sizes of all the groups. The MS_W is a measure of the intrinsic variability of the system being studied; it is the variability left over after we have explained the variations in means among the groups caused by the main effects αi.

The F-test for the presence of the main effects is

$$F_{[a-1],[\Sigma n-a]} = \frac{MS_G}{MS_W}$$

ANOVA is an immensely powerful tool that is extensible to many different situations. One of the simplest extensions of what you have already learned is to test for the presence of differences of the means of *more than two* samples at a time; however, the models can become much more sophisticated than the simple one-factor ANOVA introduced in this section. For example,

- We can compare measurements in the *same samples* with various *treatments*; for example, it is possible to use the paired-comparisons ANOVA to see if a measurement for each specific subject *before* administration of a medication is the same as the measurement for that subject after treatment. Such a comparison naturally reduces the MSwithin and increases what we call the *power* of the test.
- It is possible to look at the effects of two different factors at the same time (two-way ANOVA) and even of many factors at once (multifactorial ANOVA). For example, we could study the effects of two different doping agents on the performance of integrated circuitry. In addition to telling if the different doping agents individually change performance, the two-way ANOVA would also reveal if the effect of one factor was different as a function of the other factor. We call such dependencies *interactions*
- ANOVA with linear regression uses a model where an element of prediction based on the optimal least-squares regression coefficient to partition the variability into a component related to the regression and the residual variability assigned to the *Unexplained Mean Square Error*.

In all of these ANOVAs, the MS_W or *Unexplained Variability* or *Mean Square Error* (MS_E)tells us how much more work we have to do to understand the phenomenon we are studying; if the MS_E is large after we have explained part of the variability using models based on different classifications or regressions, then we still have mysteries to probe and to discover in the data.[94]

[94] These terms (MS_W, MS_E and so on) are used interchangeably in applied statistics; different texts and journals have their preferred usage, but we should be ready to recognize all of them, so the text here is deliberately written with several versions of the terms.

8.9 Testing for the Equality of Two Means in an Investigation of Possible Dishonesty

There are other tests for evaluating the hypothesis that the means of two samples could be as different as they are (or more different) by chance (random sampling) alone. The exact choice depends on whether we already know the parametric variances of the populations from which the samples are drawn (in which case we use the Normal distribution to describe the distribution of differences between two sample means) or if we don't know the parametric variances and have to estimate them based on the sample variances (in which case we rely on Student's-t distribution).

By far the more common class of tests involves samples with unknown parametric variances when the sample variances are comparable.

Here's an example of a situation requiring a test for equality of means.

> A computer systems administrator is investigating a possible case of insider fraud and needs to know if one of the employees has been spending a significantly longer time logged into the accounts payable database over the last three weeks compared with the previous three weeks. The reason is that the Chief Financial Officer has reported a series of unexplained discrepancies in the accounts payable files starting three weeks ago: there seem to be payments to nonexistent suppliers starting at that point. Suspicion has fallen on the Assistant Comptroller; the Chief Information Security Officer has asked for investigation of that employee's behavior. Has the suspect in fact significantly increased the length of his sessions?

The sysadmin finds the information from the log files shown in Figure 8-10.

The hypotheses are

H0: $\mu_1 \geq \mu_2$	which is sometimes expressed	H0: $\mu_1 - \mu_2 \geq 0$
H1: $\mu_1 < \mu_2$	which would be equivalent to	H1: $\mu_1 - \mu_2 < 0$

Sometimes the symbol Δ (Greek capital delta) represents the hypothesized difference between the parametric means of the two populations from which the samples are drawn. When we are testing for the equality of the parametric means, $\Delta = \mu 1 - \mu 2 = 0$.

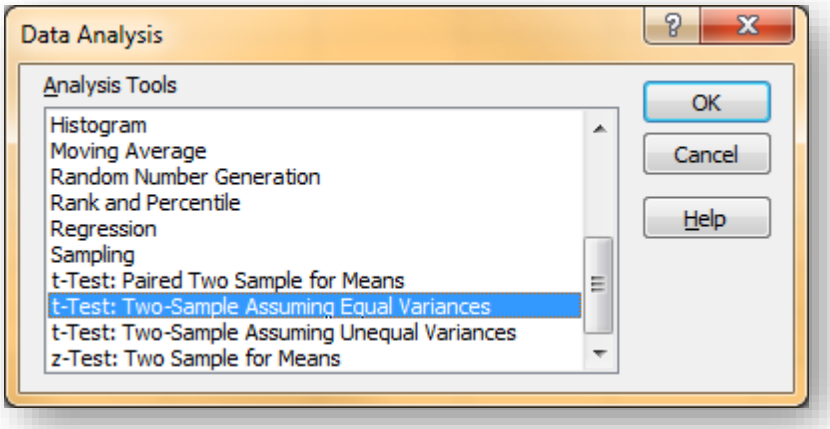
Figure 8-10. Log-file data on session length for suspected fraudster.

	A	B
1	Sample 1	Sample 2
2	18.9	30.49
3	16.89	26.63
4	15	29.69
5	22.74	25.03
6	11.58	28.56
7	20.91	29.67
8	19.04	31.29
9	17.23	24.97
10	25.69	24.86
11	18.94	25.5
12	19.29	34.33
13	21.78	33.87
14	17.58	28.55
15		32.33
16		22.7
17		23.49
18		31.05
19		29.91
20		31.26
21		29.84
22		29.7
23		29.25
24		26.91
25		30.2
26		22.68
27		22.5
28		33.74
29		30.49
30		36.11
31		24.67
32		25.54
33		30.81
34		26.9
35		27.53
36		28.7
37		31.03
38		21.99
39		25.51
40		36.7
41		32.19
42		35.49
43		30.77

8.10 T-Tests in Data Analysis

The easiest interface for a t-test is to use the Data | Data Analysis menu shown in Figure 8-11, which brings up a menu of options for t-tests.

Figure 8-11. Data Analysis tools for t-tests of means.



The last four functions shown in Figure 8-11 have the following purposes:

- t-test: Paired Two Sample for Means – used when the same subject is measured twice; e.g., to compare effect of a drug treatment by measuring blood hemoglobin concentration before and after administration of the dose on each subject rather than simply considering each group to be a random sample (discussed later in this course).
- t-test: Two Sample Assuming Equal Variances – two random samples with parametric variances thought to be identical.
- t-test: Two Sample Assuming Unequal Variances – two random samples with parametric variances thought to be different.
- z-test: Two Sample for Means – two random samples, each with a known parametric variance.

The first question is whether the observed sample variances are consistent with an assumption of homoscedasticity for the parametric variances.

The Data Analysis | Descriptive Statistics tool provides results which are shown slightly reformatted in Figure 8-12.

Figure 8-12. Descriptive statistics for two samples.

Statistic	*Sample 1*	*Sample 2*
Mean	18.890	28.891
Standard Error	0.983	0.592
Median	18.94	29.68
Mode	#N/A	30.490
Standard Deviation	3.544	3.838
Sample Variance	12.560	14.732
Kurtosis	0.928	-0.628
Skewness	-0.134	0.028
Range	14.110	14.710
Minimum	11.58	21.99
Maximum	25.69	36.70
Sum	245.57	1213.43
Count	13	42

Next is to use the F-test for equality of two variances to test for homoscedasticity using the sample variances (12.560 and 14.732). As usual, we can pick the larger variance for the numerator and the smaller for the denominator so we can use the right-sided probability function given by =F.DIST.RT.

The hypotheses for this test are

H0: $\sigma^2_1 = \sigma^2_2$ and H1: $\sigma^2_1 \neq \sigma^2_2$

Figure 8-13 shows the results. There is no reason to suspect that the parametric variances are unequal (p = 0.401ns), so we can choose the equal-variances version of the t-test for quality of means:

t-Test: Two Sample Assuming Equal Variances.

Figure 8-13. Test for equality of parametric variances in two samples.

CHECK FOR HOMOSCEDASTICITY:	Value	Formula
F-test for equal variances	1.173	=+S7/R7
P(H0) for F-test	0.401	=F.DIST.RT(L18,S14-1,R14-1)
F-test result:	ns -- variances are equal	

Figure 8-14 shows the menu and entered data for the EXCEL t-Test: Two Sample Assuming Equal Variances data-analysis routine.

The results are unequivocal:

Because our null and alternative hypotheses are asymmetric

(H0: $\mu_1 \geq \mu_2$ and H1: $\mu_1 < \mu_2$),

we use the *one-tail probability* (highlighted in bold in Figure 8-15) which is unquestionably extremely significant (1.555E11). There is virtually no question that the suspect has increased the length of his sessions significantly.

Figure 8-14. Data locations entered into menu for t-test.

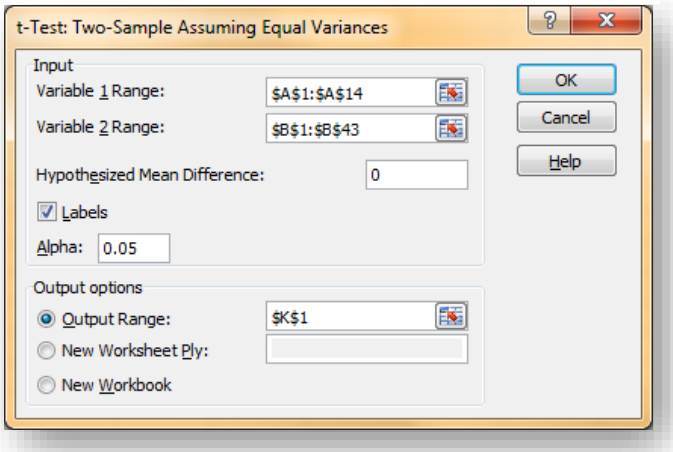


Figure 8-15. Results of t-test for equality of means assuming homoscedasticity.

t-Test: Two-Sample Assuming Equal Variances		
	Sample 1	*Sample 2*
Mean	18.89	28.89
Variance	12.56	14.73
Observations	13	42
Pooled Variance	14.24	
Hypothesized Mean Difference	0	
df	53	
t Stat	-8.350	
P(T<=t) one-tail	**1.55E-11**	
t Critical one-tail	1.674	
P(T<=t) two-tail	3.10E-11	
t Critical two-tail	2.006	

Note that if we had been testing symmetrical hypotheses (H0: $\mu_1 = \mu_2$ and H1: $\mu_1 \neq \mu_2$), the appropriate probability would have been the P(T<=t)two-tail figure which, not surprisingly, is exactly twice (3.10E-11) the figure we chose (1.55E-11). That probability includes the likelihood that we could observe a difference of this magnitude or greater *in either direction.*

8.11 Critical Values

As a note of historical interest, the data analysis routine also provides *critical values* that demarcate the boundaries of the 0.05 "Alpha" stipulated in Figure 8-14.

The t Critical one-tail value is $t_{05[53]} = 1.674$ and denotes the value of Student's t with

$$df = (n_1 + n_2 - 2) = (13 + 42 - 2) = 53$$

where 0.05 of the distribution lies to the left of this critical value. We can verify this assertion using

=T.DIST(-1.674, 53, 1) = 0.0500

as expected for the left tail. The same check could use

=T.DIST.RT(1.674,53) = 0.0500

for the right tail..

Similarly, the t Critical two-tail value represents

$$t_{025[53]} = 2.006$$

and demarcates 0.025 to the right of 2.006 and 0.025 to the left of -2.006

Again, verifying these assertions using EXCEL functions for the left and right tails, respectively,

=T.DIST(-2.006,53,1) = 0.0250

=T.DIST.RT(2.006,53) = 0.0250

These *critical values* are holdovers from a time when statisticians did not have ready access to computers and relied on printed tables of probabilities for statistical decisions. Figure 8-16 shows such a table; users would look up the proabaility for a two-tailed critical value along the top and go down the column for the critical value for degrees of freedom.[95]

Figure 8-16. Image of a statistical table from a 1981 set of tables.

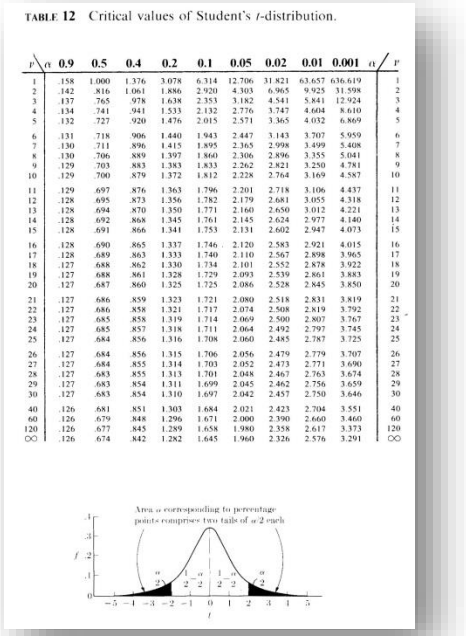


TABLE 12 Critical values of Student's *t*-distribution.

ν \ α	0.9	0.5	0.4	0.2	0.1	0.05	0.02	0.01	0.001	α / ν
1	.158	1.000	1.376	3.078	6.314	12.706	31.821	63.657	636.619	1
2	.142	.816	1.061	1.886	2.920	4.303	6.965	9.925	31.598	2
3	.137	.765	.978	1.638	2.353	3.182	4.541	5.841	12.924	3
4	.134	.741	.941	1.533	2.132	2.776	3.747	4.604	8.610	4
5	.132	.727	.920	1.476	2.015	2.571	3.365	4.032	6.869	5
6	.131	.718	.906	1.440	1.943	2.447	3.143	3.707	5.959	6
7	.130	.711	.896	1.415	1.895	2.365	2.998	3.499	5.408	7
8	.130	.706	.889	1.397	1.860	2.306	2.896	3.355	5.041	8
9	.129	.703	.883	1.383	1.833	2.262	2.821	3.250	4.781	9
10	.129	.700	.879	1.372	1.812	2.228	2.764	3.169	4.587	10
11	.129	.697	.876	1.363	1.796	2.201	2.718	3.106	4.437	11
12	.128	.695	.873	1.356	1.782	2.179	2.681	3.055	4.318	12
13	.128	.694	.870	1.350	1.771	2.160	2.650	3.012	4.221	13
14	.128	.692	.868	1.345	1.761	2.145	2.624	2.977	4.140	14
15	.128	.691	.866	1.341	1.753	2.131	2.602	2.947	4.073	15
16	.128	.690	.865	1.337	1.746	2.120	2.583	2.921	4.015	16
17	.128	.689	.863	1.333	1.740	2.110	2.567	2.898	3.965	17
18	.127	.688	.862	1.330	1.734	2.101	2.552	2.878	3.922	18
19	.127	.688	.861	1.328	1.729	2.093	2.539	2.861	3.883	19
20	.127	.687	.860	1.325	1.725	2.086	2.528	2.845	3.850	20
21	.127	.686	.859	1.323	1.721	2.080	2.518	2.831	3.819	21
22	.127	.686	.858	1.321	1.717	2.074	2.508	2.819	3.792	22
23	.127	.685	.858	1.319	1.714	2.069	2.500	2.807	3.767	23
24	.127	.685	.857	1.318	1.711	2.064	2.492	2.797	3.745	24
25	.127	.684	.856	1.316	1.708	2.060	2.485	2.787	3.725	25
26	.127	.684	.856	1.315	1.706	2.056	2.479	2.779	3.707	26
27	.127	.684	.855	1.314	1.703	2.052	2.473	2.771	3.690	27
28	.127	.683	.855	1.313	1.701	2.048	2.467	2.763	3.674	28
29	.127	.683	.854	1.311	1.699	2.045	2.462	2.756	3.659	29
30	.127	.683	.854	1.310	1.697	2.042	2.457	2.750	3.646	30
40	.126	.681	.851	1.303	1.684	2.021	2.423	2.704	3.551	40
60	.126	.679	.848	1.296	1.671	2.000	2.390	2.660	3.460	60
120	.126	.677	.845	1.289	1.658	1.980	2.358	2.617	3.373	120
∞	.126	.674	.842	1.282	1.645	1.960	2.326	2.576	3.291	∞

Today, few statisticians use statistical tables unless they also like to use tables of logarithms, or to perform statistical calculations by hand or using a calculator, or to use overhead projectors and photocopies on acetates for prepared images in lectures instead of computer programs and LCD projectors, or to chisel cuneiform symbols on clay tablets instead of using email.

However, if you are stranded on Pluto without access to your cerebral computational brain implant, having a set of statistical tables may allow you to check a published calculation or two.[96]

[95] (Rohlf and Sokal 1981) p 81. Used with kind permission of author F. J. Rohlf.

[96] Geeks should note that carrying statistical tables around will *not* increase your attractiveness to potential dates.

8.12 ANOVA: Single Factor vs T-test for Equality of Means

The Student's t-test for equality of means produces results consistent with those of an ANOVA. To demonstrate this similarity, Figure 8-17 shows the ANOVA using EXCEL's Data Analysis routine Anova: Single Factor for the same data as the t-test just discussed.

Students should note the equality of the

Figure 8-18. ANOVA for same data as t-test example.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Sample 1	13	245.57	18.89	12.56
Sample 2	42	1213.43	28.89	14.73

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	992.9637	1	992.9637	69.7	3.10E-11	4.02
Within Groups	754.7354	53	14.2403			
Total	1747.699091	54				

P(T<=t) two-tail = 3.10E-11 from the t-test and the

P-value = 3.10E-11 in the ANOVA for the F[1,53] = 69.7***.

ANOVA of this kind always performs a *two-tailed* F-test for the H0 of *equality* of means.

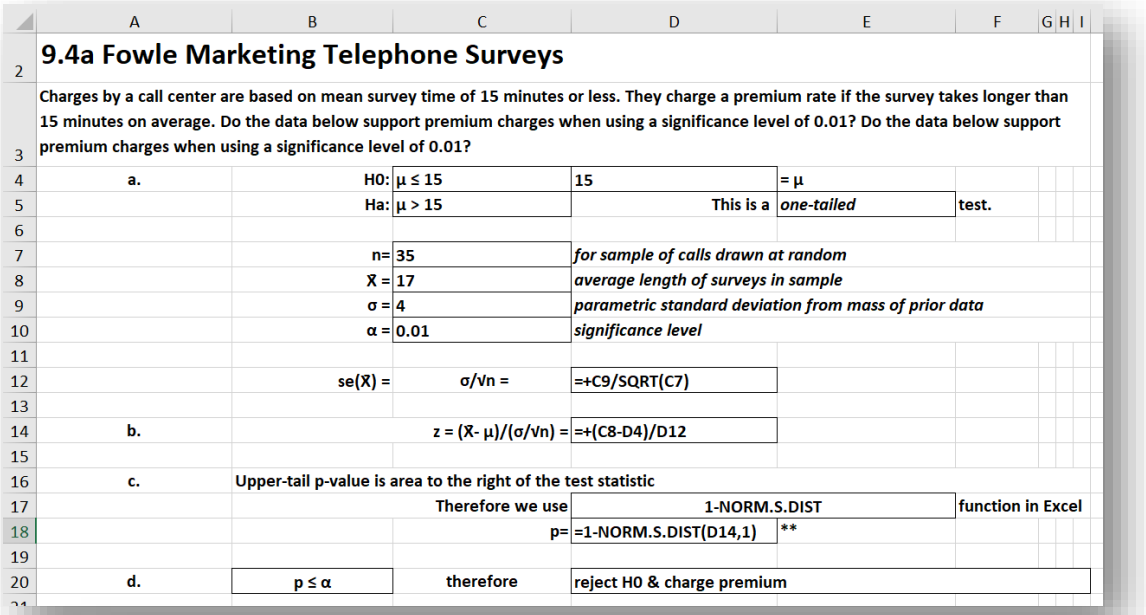
In the t-test and in the ANOVA, the null hypothesis is rejected: the difference between the means is extremely statistically significant at the 0.001 level (there is essentially no realistic chance of encountering a difference this large or larger by chance alone if the null hypothesis of equality is true) and therefore we accept the alternate hypotheses:

- In the original t-test shown in § 8.10, that the mean session length is *longer* in the second sample than in the first sample (a *one*-tailed test);
- In the ANOVA, that the mean session length is *different* between the samples (a *two*-tailed test). Given that the observed difference is that $\overline{Y}_2 > \overline{Y}_1$, we come to the same conclusion as in the t-test: the Assistant Comptroller has indeed been spending a lot more time logged into the accounts payable database in these last three weeks than in the three weeks before. Maybe it's time to do some more forensic research on the case.

8.13 Testing for Equality of Means Given Parametric Mean and Parametric Standard Deviation v Sample Mean

Sometimes we have such a large amount of historical information about the mean and standard deviation of a measurement that we are willing to consider those *parametric* statistics. In §5.7, we discussed how to compute *confidence limits* based on parametric mean and parametric standard deviation. The same calculations can be adapted to test the hypothesis that and observed *sample* mean and observed *sample* standard deviation are consistent with the null hypothesis. Here are screenshots of the formulas and the calculated values from a homework exercise in the QM213 course at Norwich University.

Figure 8-19. Using parametric mean & standard deviation for an hypothesis test about a sample mean.



	A	B	C	D	E	F
2	**9.4a Fowle Marketing Telephone Surveys**					
3	**Charges by a call center are based on mean survey time of 15 minutes or less. They charge a premium rate if the survey takes longer than 15 minutes on average. Do the data below support premium charges when using a significance level of 0.01? Do the data below support premium charges when using a significance level of 0.01?**					
4	**a.**	**H0:**	**μ ≤ 15**	**15**	**= μ**	
5		**Ha:**	**μ > 15**	**This is a**	***one-tailed***	**test.**
6						
7		**n=**	**35**	***for sample of calls drawn at random***		
8		**X̄ =**	**17**	***average length of surveys in sample***		
9		**σ =**	**4**	***parametric standard deviation from mass of prior data***		
10		**α =**	**0.01**	***significance level***		
11						
12		**se(X̄) =**	**σ/√n =**	**=+C9/SQRT(C7)**		
13						
14	**b.**		**z = (X̄- μ)/(σ/√n) =**	**=+(C8-D4)/D12**		
15						
16	**c.**	**Upper-tail p-value is area to the right of the test statistic**				
17			**Therefore we use**	**1-NORM.S.DIST**	**function in Excel**	
18			**p=**	**=1-NORM.S.DIST(D14,1)**	**	
19						
20	**d.**	**p ≤ α**	**therefore**	**reject H0 & charge premium**		

9.4a Fowle Marketing Telephone Surveys

Charges by a call center are based on mean survey time of 15 minutes or less. They charge a premium rate if the survey takes longer than 15 minutes on average. Do the data below support premium charges when using a significance level of 0.01? Do the data below support premium charges when using a significance level of 0.01?

a.	**H0:**	**μ ≤ 15**	**15**	**= μ**
	Ha:	**μ > 15**	**This is a**	***one-tailed*** **test.**
	n=	**35**	***for sample of calls drawn at random***	
	X̄ =	**17**	***average length of surveys in sample***	
	σ =	**4**	***parametric standard deviation from mass of prior data***	
	α =	**0.01**	***significance level***	
	se(X̄) =	**σ/√n =**	**0.6761**	
b.	**z = (X̄- μ)/(σ/√n) =**		**2.9580**	
c.	**Upper-tail p-value is area to the right of the test statistic**			
	Therefore we use		**1-NORM.S.DIST**	**function in Excel**
	p=		**0.001548**	**
d.	**p ≤ α**	**therefore**	**reject H0 & charge premium**	

8.14 Computing a t-test for Equality of Means without Raw Data

Finally, in case you need it, here is the intimidating general formula for the t-test for equality of two means if you *don't* have acccess to the raw data but only the means, the sample variances, and the sample sizes. The symbol Δ is just the hypothesized difference between the parametric means – usually 0.

$$t_{df} = \frac{(\bar{Y}_1 - \bar{Y}_2 - \Delta)}{\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}} \quad \text{where} \quad df = \frac{\left(\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s^2_1}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s^2_2}{n_2}\right)^2}$$

Aren't you glad you use EXCEL? Although there is no instant function for these calculations, as long as you label all the steps of your calculations, you will have no problem computing the values step by step and then using the =T.DIST.2T function to compute the probability of observed t-value if the null hypothesis were true.

8.15 The T.TEST Function

For completeness, let's look at another t-test for the equality of means, this time using raw data and the EXCEL =T.TEST function, which requires the array (both columns of data but NOT the headings), the number of tails (1 or 2), and the choice of model:

- **Type = 1 for paired comparisons** in which the same entities are observed under two different circumstances; e.g., if the same glue job were repeated with standard application glue and then with molecular transubstantiation glue. There must, naturally, be exactly the same number of observations if we are doing a paired comparison. Another typical example is a before/after comparison of individual test subjects; e.g., a paired comparison of *each person's* weight before and after a month of a special diet.
- **Type = 2 for two samples** (as in our example of the possibly corrupt Assistant Comptroller) where we assume that the parametric variances are equal (homoscedasticity). In our case, lacking any reason to do otherwise, we choose this assumption.
- **Type = 3 for two heteroscedastic samples** (where we assume that the parametric variances are unequal).

Using the same data as in the t-test of § 8.9 *et seq.*,

=T.TEST(A2:A14,B2:B43,1,2) = 1.55E-11

Which is exactly the same as the result of the one-tailed t-test in §8.10 (Figure 8-15).

Similarly,

=T.TEST(A2:A14,B2:B43,2,2) = 3.10E-11

exactly as in both the two-tailed t-test and in the ANOVA of §8.12 (and Figure 8-17).

