Which Test is This?

In this class we will encounter many different hypothesis tests. In a traditional freshman statistics course we might see around eight such tests. In an engineering statistics course, we may encounter a dozen or more, along with the related confidence intervals. Most of these tests can be performed in our calculator (TI-84). The goal of this handout is to give some key things to look for each type of hypothesis test. It's not designed to help you set them up or interpret them, but the first battle is to know which test you need to conduct for a given problem.

I. Z-Test for Means (Hypothesis Test for One Variable)

- Is there only one variable? Check the data table to be sure the any tables are displaying a single data set, or multiple data sets, or related variables. If it really is only one, then this might work. Also check that you are testing means and not proportions.
- Is σ known? If it is, this test might be okay. Check the sample size.
- If σ is known and the sample size $n \geq 30$, **and** the data is clearly and strongly normally distributed, this may be okay. A sample size of $n \geq 40$ is generally better. Larger sample sizes, like $n \geq 50$, you can use this test with s, but if it's not given in the problem you will have to calculate it separately. If σ is not known, generally speaking, it is safer to use a T-test instead. If you need accuracy, use with sample sizes $n \geq 40$. If there is any question about normality, or any uncertainty at all, use the T-test anyway.
- Generally speaking, use this test sparingly if there is any doubt about the conditions at all. We learn this one first because the concepts around it are well-known, but real world circumstances rarely satisfy it.

II. T-Test for Means (Hypothesis Test for One Variable)

- Must be done with only one variable. See the same checks suggested for the data set for the Z-Test. Check that you are testing means and not proportions.
- Is σ known? If the answer is no, it's generally best to use a T-Test.
- Is the sample size $n < 40$? If it is, this is the right test whether σ is known or not.
- If the data set is approximately normal, the T-Test is generally good. The larger the sample size the greater the deviation from normal (or the greater the uncertainty about the normality) is acceptable.

III. Z-Test for Proportions

- Proportions can be given as a count with a sample size, or a sample size and a percent. But not all percentages are proportions. For instance, if alcohol content of wine is being measured (in % alcohol by volume), this is a means test (either T or Z). If we are conducting a survey and 53% of a sample of 2600 responded in a particular way, this is a proportion, and not a mean.
- There is no T-Test for proportions. If the sample size is too small, one can use a binomial distribution. However, this is ***extremely*** atypical.
- Be sure you are only comparing one set of data, typically compared to background assumptions. If there are two samples being compared, you need to two-sample test.

- Recall that $np > 10, nq > 10$. (Though you may be able to get away with 5 if you don't need high accuracy.

IV. T-Test for Two Means (Dependent) – Test on the Differences

- Are there two samples? If so, we need to ask a couple more questions.
- Are the sample sizes the same? If the answer is no, this is not the test you want.
- If the sample sizes are the same, is the data displayed in related pairs? If so, check that both sets of data are measuring the same thing under different circumstances. If not, this is a correlation/regression test, and not this one.
- If the data is in pairs, check the problem to ensure that the samples are related measure by measure: either taken on the same object at two different times, or taken at the same time with two different methods.
- If this test does apply, you must take the difference of each pair, and do a (one sample) T-Test on those differences. This is not the regular Two-Sample T-Test.
- There is not really a Z-Test for the dependent case (though you can use it if the sample size is extremely large) since we are unlikely to know the standard deviation of the difference *a priori*.
- For this and all two-sample tests, the order of subtraction is not that important for two-tailed tests, but matters a great deal for one-tailed tests. It's generally better to think about which order makes the most sense rather than doing the default order all the time.

V. Two-Sample T-Test (Independent)

- Are there two samples? If there are more than two samples, this is not the right test.
- If the sample sizes are different, this probably is the right test.
- If the sample sizes are the same, check that the data is not dependent (see §IV first).
- Is σ known? If it is, and the sample sizes are large, you can use the Two-Sample Z-Test, but if the answer to one of these questions is "no" (ie. The sample size is small and/or σ is not known), then this is the correct test.
- For this and all two-sample tests, the order of subtraction is not that important for two-tailed tests, but matters a great deal for one-tailed tests. It's generally better to think about which order makes the most sense rather than doing the default order all the time.

VI. Pooled Two-Sample T-Test (Independent)

- This test used to be more popular, but it has fallen out of favour. In general, assume it does not apply if there is any question.
- The main assumption here is that the two samples have identical standard deviations (the behaviour is the same, only the mean is shifted). This is rarely a safe assumption, so use it exceedingly sparingly.

VII. Two-Sample Z-Test (Independent)

- This test, like one-sample Z-Test, does exist, but generally speaking the standard deviation for both data sets must be known and the sample sizes must be large. The T-Test is a better default for most of the situations we are likely to encounter. See the caveats for the one-sample Z-Test if you are unsure whether the two-sample Z-Test or the T-Test applies.

- For this and all two-sample tests, the order of subtraction is not that important for two-tailed tests, but matters a great deal for one-tailed tests. It's generally better to think about which order makes the most sense rather than doing the default order all the time.

VIII. Two-Proportion Z-Test

- Check that there are two samples employing proportions. To tell the difference between means and proportions, compare with steps for one-proportion test.
- If there are three or more proportions, this is probably a χ^2 test or some other test.
- This test cannot handle very small sample sizes. Be sure to check for the acceptability of the normality conditions.

IX. One-Way ANOVA

- Do you have three or more samples? Then, this is probably the correct test.
- The assumption for the null hypothesis for this test is that all the means are identical. The alternative is that one (or more) mean(s) is (are) different than the one (others).
- There is no restriction on the data set sizes here. They are generally small, but can be used with larger ones just as well.
- If the null is rejected, then one can apply Tukey's Procedure to separate out which are alike and which are not. If the null is not rejected, then Tukey's Procedure does not apply.

X. Correlation/Slope of the Regression Line T-Test

- Whether you are testing the correlation or the slope, our calculator treats this as the same test, and they use the same test statistic.
- In the default case, we are testing against zero, in which case the units are not important. If you are comparing the slope to a specific value, then you do need to concern yourself with which formula you are using.
- The related confidence interval in the calculator uses the slope and not the correlation. You can use the confidence interval to conduct your test if the comparison is to a non-zero slope.
- A T-Test is almost always used regardless of the sample sizes. It is certainly possible that the sample size is sufficiently large to justify using a Z-Test version of this, however, since σ is also never known in advance, and our data sets generally won't be that large, stick with the T-Test.
- Do check for approximate normality with a diagnostic plot.
- The same sort of test can be conducted on any of the coefficients we find in any type of regression model on individual parameters, but these must be conducted by hand using the general formula.

XI. Model Utility Test

- This test is used to determine if our model overall is a good fit for the data, not just a particular coefficient to be non-zero.
- The test statistic depends on the R^2 value and produces an F-statistic like the ANOVAs. And like ANOVAs, the test is very general. The null hypothesis is that all the coefficients are zero, and the alternative hypothesis is that at least one is non-zero. To test specific parameters in the model, you need the T-Test above on each parameter.

XII. Goodness of Fit Tests (for Proportions)

- First, determine if there is a model you are attempting to fit (that's it binomial, or proportions are specified, etc.), or if the only model is that all proportions are the same. This affects the probabilities that are used to calculate the values for "expected", and so your results will not be meaningful if you use the wrong ones.
- These can be done in the calculator, but you have to calculate the "expected" values yourself.
- Use this when data is displayed in a table with a single row. If the table has more than one row, go to the Two-Way Contingency Table/Test for Independence/Homogeneity test below.

XIII. Test of Independence/Test of Homogeneity

- Is your data displayed in two or more rows, and 2 or more columns? Then this test probably applies if we are talking about counts/proportions. If either the rows or the columns are less than one, you need the Goodness of Fit Test.
- The test of independence and test of homogeneity are basically the same, only the way we state the hypothesis is different. Unlike some other tests, the null/alternative hypotheses are easier stated in words than in algebraic statements.
- The contingency tables may be familiar from our work with conditional probabilities. This test essentially tests whether $P(A) = P(A|B)$.

This is by no means all the hypothesis tests there are, but these are some of the big ones. The first step in conducting any test, is knowing which test is appropriate.

Practice Problems.

For each of the problems below, determine which test is appropriate for the scenario described and set up the appropriate null and alternative hypotheses. You do not need to conduct the test. Some tests may be used more than one. Some tests may not appear at all.

1. The alternating current (AC) breakdown voltage of an insulating liquid indicates its dielectric strength. An article gave the accompanying sample observations on breakdown voltage (kV) of a particular certain under certain conditions. Previous investigations suggested that the mean of breakdown voltage is 55 kV.

62	**50**	**53**	**57**	**41**	**53**	**55**	**61**	**59**	**64**
50	53	64	62	50	68	54	55	57	50
55	50	56	55	46	55	53	54	52	47
47	55	57	48	63	57	57	55	53	59
53	52	50	55	60	50	56	58		

2. The Pew Forum on Religion and Public Life reported in December 2009 that in a survey of 2003 American adults, 25% said they believed in astrology. The average of the previous 10 decades of surveys was 23%. Does this data represent a significant increase from previous years?
3. A study of the ability of individuals to walk in a straight line reported the accompanying data on cadence (strides per second) for a sample of $n = 20$ randomly selected healthy men.

0.95	**0.78**	**0.85**	**0.93**	**0.92**	**0.93**	**0.95**	**1.05**	**0.93**	**0.93**
0.86	1.06	1.00	1.06	0.92	0.96	0.85	0.81	0.81	0.96

A normal probability plot gives substantial support to the assumption that the population distribution of cadence is approximately normal. Determine if the mean is significantly different from 1 stride per second.

4. The calibration of a scale is to be checked by weighing a 10 kg test specimen 25 times. Suppose the results of different weighings are independent of one another and that the weight on each trial is normally distributed with $\sigma = 0.200\ kg$. Let μ denote the true average weight reading of the scale. Suppose a sample this time determined the mean was 10.1 kg.
5. Hexavalent chromium has been identified as an inhalation carcinogen and an air toxin of concern in a number of different locales. An article gave the accompanying data on both indoor and outdoor concentrations (nanograms/m^2) for a sample of houses selected from a certain region. Determine if the indoor and outdoor measurements are different.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Indoor	0.07	0.08	0.09	0.12	0.12	0.12	0.13	0.14	0.15
Outdoor	0.29	0.68	0.47	0.54	0.97	0.35	0.49	0.84	0.86
	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Indoor	0.15	0.17	0.17	0.18	0.18	0.18	0.18	0.19	0.20
Outdoor	0.28	0.32	0.32	1.55	0.66	0.29	0.21	1.02	1.59
	19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
Indoor	0.22	0.22	0.23	0.23	0.25	0.26	0.28	0.28	0.29
Outdoor	0.90	0.52	0.12	0.54	0.88	0.49	1.24	0.48	0.27
	28	**29**	**30**	**31**	**32**	**33**			
Indoor	0.34	0.39	0.40	0.45	0.54	0.62			
Outdoor	0.37	1.26	0.70	0.76	0.99	0.36			

6. In an experiment to compare bearing strengths of pegs inserted in two different types of mounts, a sample of 14 observations on stress limit for red oak mounts resulted in a sample mean and sample standard deviation of 8.48 MPa and 0.79 MPa respectively, whereas a sample of 12 observations when Douglas fir mounts were used gave a mean of 9.36 MPa and a standard deviation of 1.52. Test whether or not true average stress limits are identical for the two types of mounts.
7. An article reports the following data on total Fe (iron) for four types of iron formation (1=carbonate, 2=silicate, 3=magnetite, 4=hematite). Carry out an analysis of variance F test.

1.	**20.5**	**28.1**	**27.8**	**27.0**	**28.0**
	25.2	25.3	27.1	20.5	31.3
2.	26.3	24.0	26.2	20.2	23.7
	34.0	17.1	26.8	23.7	24.9
3.	29.5	34.0	27.5	29.4	27.9
	26.2	29.9	29.5	30.0	35.6
4.	36.5	44.2	34.1	30.3	31.4
	33.1	34.1	32.9	36.3	25.5

The efficiency ratio for a steel specimen immersed in a phosphating tank is the weight of the phosphate coating divided by the metal loss. An article gave the accompanying data on tank temperature (x) and efficiency ratio (y). Construct a regression line and determine if the correlation is non-zero.

Temperature	**170**	**172**	**173**	**174**	**175**	**176**	**177**	**180**	**180**
Efficiency Ratio	0.84	1.31	1.42	1.03	1.07	1.08	1.04	1.80	1.45
Temperature	**180**	**180**	**180**	**181**	**181**	**182**	**182**	**182**	**182**
Efficiency Ratio	1.61	2.13	2.15	0.84	1.43	0.90	1.81	1.94	2.68
Temperature	**184**	**184**	**185**	**186**	**188**				
Efficiency Ratio	1.49	2.52	3.00	1.87	3.08				

8. Astringency is the quality in a wine that makes the wine drinker's mouth feel slightly rough, dry and puckery. A paper reported on an investigation to assess the relationship between perceived astringency and tannin concentration using various analytic methods. Here is data provided by the authors on x=tannin concentration by protein precipitation and y=perceived astringency as determined by a panel of tasters.

x	**0.718**	**0.808**	**0.924**	**1.000**	**0.667**	**0.529**	**0.514**	**0.559**	**0.766**
y	0.428	0.480	0.493	0.978	0.318	0.298	-0.224	0.198	0.326
x	0.470	0.726	0.762	0.666	0.562	0.378	0.779	0.674	0.858
y	-0.336	0.765	0.190	0.066	-0.221	-0.898	0.836	0.126	0.305
x	0.406	0.927	0.311	0.319	0.518	0.687	0.907	0.638	0.234
y	-0.577	0.779	-0.707	-0.610	-0.648	-0.145	1.007	-0.090	-1.132
x	0.781	0.326	0.433	0.319	0.238				
y	0.538	-1.098	-0.581	-0.862	-0.551				

Relevant summary quantities are as follows: $\sum x_i = 19.404, \sum y_i = -0.549, \sum x_i^2 = 13.248032, \sum y_i^2 = 11.835795, \sum x_i y_i = 3.497811, S_{xx} = 1.48193150, S_{yy} = 11.82637622, S_{xy} = 3.83071088$. Use this information to determine if a linear model is appropriate, and if so, is the slope of the regression line negative?

9. The following data on mass rate of burning x and flame length y is representative of that which appeared in an article. Construct a power function model and conduct an appropriate test of the model.

x	**1.7**	**2.2**	**2.3**	**2.6**	**2.7**	**3.0**	**3.2**
y	1.3	1.8	1.6	2.0	2.1	2.2	3.0
x	3.3	4.1	4.3	4.6	5.7	6.1	
y	2.6	4.1	3.7	5.0	5.8	5.3	

10. In a genetics experiment, investigators looked at 300 chromosomes of a particular type and counted the number of sister-chromatid exchanges on each. A Poisson model was hypothesized for the distribution to the data by first estimating μ and then combining the counts for $x = 8, x = 9$ into one cell.

x = number of exchanges	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Observed Counts	6	24	42	59	62	44	41	14	6	2

11. Does the phase of the moon have any bearing on birthrate? Each of the 222,784 births the occurred during a period encompassing 24 full lunar cycles was classified according to lunar phase. The following data is consistent with summary quantities that appeared in an article on the topic.

Lunar Phase	**# of Days in Phase**	**# of Births**
New Moon	24	7860
Waxing Crescent	152	48,442
First Quarter	24	7579
Waxing Gibbous	149	47,814
Full Moon	24	7711
Waning Gibbous	150	47,595
Last Quarter	24	7733
Waning Crescent	152	48,230

State and test the appropriate hypotheses to answer the question posed in this exercise.

12. The accompanying data on degree of spirituality for samples of natural and social scientists at research universities as well as for a sample of non-academics with graduate degrees. Is there substantial evidence for concluding that the three types of individuals are not homogeneous with respect to their degree of spirituality? State the appropriate hypotheses.

Degree of Spirituality

	Very	Moderate	Slightly	Not at all
Natural Science	56	162	198	211
Social Science	56	223	243	239
Graduate Degree	109	164	74	28

13. Scientists think that robots will play a crucial role in factories in the next several decades. Suppose that in an experiment to determine whether the use of robots to weave computer cables is feasible, a robot was used to assemble 500 cables. The cables were examined and there were 15 defectives. If human assemblers have a defect rate of 0.035, does this data support the hypothesis that the proportion of defectives is lower for robots than for humans?
14. The accompanying data on cube compressive strength (MPa) of concrete specimens appeared in an article in 2009. Determine if the true value of cube compressive strength is less than 100 MPa.

112.3	**97.0**	**92.7**	**86.0**	**102.0**	**99.2**	**95.8**	**103.5**	**89.0**	**86.7**

15. The National Health Statistics Reports dates Oct. 22, 2008 included the following information on the heights in inches for non-Hispanic white females. Is there sufficient evidence to think that older women are more than an inch shorter than younger women?

Age	Sample Size	Sample Mean	Standard Error Mean
20-39	866	64.9	0.09
60+	934	63.1	0.11

16. Tensile strength tests were carried out on two different grades of wire rod, resulting in the accompanying data. Does the data provide compelling evidence for concluding that true average strength for the 1078 grade exceeds that for the 1064 grade by more than 10 $\frac{kg}{mm^2}$?

Grade	Sample Size	Sample Mean $\left(\frac{kg}{mm^2}\right)$	Sample Standard Deviation
AISI 1064	$m = 129$	$\bar{x} = 107.6$	$s_1 = 1.3$
AISI 1078	$n = 129$	$\bar{y} = 123.6$	$s_2 = 2.0$

17. It is common practice in many countries to destroy (shred) refrigerators at the end of their useful lives. In this process material from insulating foam may be released into the atmosphere. An article gave the following data produced by 4 different manufacturers.

1.	**30.4, 29.2**	**2.**	**27.7, 27.1**
3.	27.1, 24.8	4.	25.5, 28.8

Does it appear that true average foam density is not the same for all these manufacturer? Carry out an appropriate test of hypotheses.

18. Hydrogen content is conjectured to be an important factor in porosity of aluminum alloy castings. An article gives the accompanying data on x=content and y=gas porosity for one particular measurement technique. Test at level 0.05 to see whether the population correlation coefficient differs from zero.

x	**0.18**	**0.20**	**0.21**	**0.21**	**0.21**	**0.22**	**0.23**
y	0.46	0.70	0.41	0.45	0.55	0.44	0.24
x	0.23	0.24	0.24	0.25	0.28	0.30	0.37
y	0.47	0.22	0.80	0.88	0.70	0.72	0.75

19. Criminologists have long debated whether there is a relationship between weather conditions and the incidence of violent crime. An article classified 1361 homicides according to season, resulting in the accompanying data. Test the null hypothesis of equal proportions using $\alpha = 0.01$ by using the chi-squared table to say as much as possible about the P-value.

Winter	**Spring**	**Summer**	**Fall**
328	334	372	327

20. Consider a large population of families in which each family has exactly three children. If the genders of the three children in any family are independent of one another, then number of male children in a randomly selected family will have a binomial distribution based on three trials. Suppose a random sample of 160 families yields the following results. Test the relevant hypotheses.

Number of Male Children	**0**	**1**	**2**	**3**
Frequency	14	66	64	16

21. Each individual in a random sample of high school students and college students was cross-classified with respect to both political views and marijuana usage, resulting in the data displayed in the accompanying two-way table. Does the data support the hypothesis that political views and marijuana usage level are independent within the population? Test the appropriate hypotheses using the level of significance 0.01.

		Usage Level		
		Never	Rarely	Frequently
Political Views	Liberal	479	173	119
	Conservative	214	47	15
	Other	172	45	85