Chapter 1: Introduction and Background

1.1 Introduction

Predicated upon the requirement of repeatability, the scientific process – in which theories are proposed, tested, updated and retested – often yields huge amounts of data, and statistical methods have been created and continue to be created to turn these data into useful conclusions and decisions, and subsequently into policy. As such, statistical techniques help decide which differences are real/significant and which are simply due to random variability. Unfortunately, statistics is a field in which statistical techniques can very easily be mis-applied often with little or no indication of such an error. Further, the dogma or core of statistical methodology knowledge is constantly evolving, ever pressing the practicing statistician to "stay current". Throughout this book we stress the proper matching of the appropriate tool (i.e., the statistical technique) with the given dataset, structure, and/or situation.

In this chapter, we review basic statistical methods covered in an introductory course in statistical methods. Topics covered in such course and reviewed here include the following: summary measures of data, basic probability results (including the binomial and normal distributions), fundamentals of sampling including the central limit theorem, one- and two-sample estimation (point estimates and confidence intervals) and hypothesis testing (for means and proportions and differences of means and of proportions), methods for categorical data analysis, one-way analysis of variance (ANOVA) including various multiple comparison procedures including Fisher's LSD, Tukey's HSD, Newman-Keuls and Bonferroni methods, and linear correlation and simple linear regression.

1.2 Random Variables and Sampling Theory

A random variable is a function resulting from a random process such as the tossing of a coin or the taking of a random sample from a given population. We denote random variables by symbols such as X, Y or Z; outcomes of these processes are denoted by the lower-case letters x, y or z. For most practical purposes, random variables can be thought of as either quantitative or qualitative. Additionally, quantitative random variables are further divided into either discrete or continuous variables; qualitative random variables are further divided into either nominal or ordinal variables. Examples of random variables are:

(a) the number of head's obtained in tossing a coin 20 times or the number of tosses required to observe the first head (discrete);
(b) a person's exact age or weight (without rounding off; continuous);
(c) a friend's hair color or blood group (nominal);
(d) a patient's pain relief self-assessment after taking a drug when chosen from the list: complete, substantial, some, and none (ordinal).

Some random variables have certain key known distributions. Examples of discrete distributions include

- the binomial (e.g. the number of head's observed in 20 tosses of a fair coin)
- the Poisson (e.g., the number of fish caught from Skaneateles Lake in a given year)
- the discrete uniform (e.g., the number of dots appearing on the face of a die after a single toss of the die)
- the geometric (e.g., the number of tosses of a coin needed until the first head is observed).

Examples of continuous distributions include

- the normal (sometimes called the Gaussian)
- the χ^2 (chi-square)
- the t
- the F
- the continuous uniform.

Some of these continuous distributions are the (null) distribution of a given test statistic, and may be first encountered when performing hypothesis tests. As a result, most basic statistics textbooks contain probability tables for some of these distributions. To illustrate, in testing for the equality of two normal population means when the variances are assumed to be equal, the two-sample test statistic has the t-distribution when the null hypothesis of equal means is true; as a result, cumulative t probabilities or quantiles are given in most statistics textbooks.

Many distributions have associated probability mass functions (for discrete) or probability density functions (for continuous). For example, for the binomial distribution with n independent trials and with success probability π, for y = 0, 1, 2, … n, the probability mass function is

$$P(Y = y) = P(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

Other probability functions can be found in a textbook on basic probability theory such as Wackerly, Mendenhall, & Scheaffer (2008). Random variables also have associated moments; the first moment is the mean (denoted μ) and the second centralized moment is the variance (denoted σ^2). Using "E( )" to denote the mathematical expected value, the respective formulas for the mean and variance are $\mu = E(Y)$ and $\sigma^2 = VAR(Y) = E[(Y - \mu)^2]$. Expected values usually involve summations or integration, and when the result is infinite, we say that the corresponding moment does not exist. For the binomial distribution, the general formulas simplify giving $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$. It is emphasized, however, that for other distributions these formulas do not hold, and in these cases, the first and second moments are found using the general definitions.

In a population of size N, a random sample of size n (n < N) implies that the probability associated with selecting any set of n population elements is the same, and is the same as the probability of selecting any other set of n element. If a random sample of size n is chosen from a known probability distribution, the field of sampling theory is concerned with characterizing and understanding what to expect for the distribution of a (derived) statistic such as the sample proportion (denoted p) or the sample mean (denoted $\bar{y}$). The Central Limit Theorem demonstrates that as long as the sample size is large enough, these sampling distributions will be approximately bell-shaped (i.e., normal in shape) even when the underlying population distribution is not normally distributed. Regardless of the sample size and the shape of the underlying population distribution, for a continuous random variable Y with mean μ and variance σ^2, it is the case that $E(\overline{Y}) = \mu$ and $VAR(\overline{Y}) = \frac{\sigma^2}{n}$. This implies that the distribution of $\overline{Y}$ is also centered at μ, and that the variability in the $\overline{Y}$ distribution can be decreased by choosing a large enough sample size.

1.3 Statistical Methods for a Single Sample

It is important to point out that sampling theory discussed in the previous paragraph is largely useful only in a theoretical sense: the practitioner rarely knows the underlying parent population or the associated population parameters. Rather, it is more common to have some data from a random sample and one wishes to use that data to estimate one or several population parameters. This can be achieved by using either a point estimate or an interval (or region) estimate and/or by using a statistical hypothesis test regarding the parameter(s). Indeed, a large part of statistical methodology is concerned

with both estimation and hypothesis testing. It's important to point out that the process of moving from a known population to a single sample (sampling theory) involves deductive logic and is inherent in the field of mathematics; the field of statistics, on the other hand, uses inductive reasoning to generalize from one's data to the larger population using inference.

To illustrate estimation and testing methods, consider taking a random sample from a population and recording a quantitative random variable. We can set a confidence interval for the true mean (μ) by using the usual t-distribution; similarly we can test whether this unknown population mean is equal to a specific number (μ_0) again using the same t- test statistic. This test statistic has the form $t = \frac{\bar{y}-\mu_0}{s/\sqrt{n}}$, where $\bar{y}$ and s are the observed sample mean and standard deviation. When the underlying parent population is normally distributed or when the sample size is sufficiently large, we use the t distribution with n-1 degrees of freedom to perform the hypothesis test. An important application of these methods involves the paired t-test – such as is useful for a "before-and-after" study or for a "twins study" – for example in which one sibling gets drug A and the other gets drug B. In this case, we work with the differences of the quantitative measurements, finding the average ($\bar{d}$) and standard deviation (s_d) of these differences, and using the analogous test statistic $t = \frac{\bar{d}}{s_d/\sqrt{n}}$. A third illustration for a single sample involves the binomial setup, where we are sometimes interested in setting a confidence interval for the unknown success probability (π). Instead of using the usual estimator, $p = \frac{y}{n}$, for finding a 95% confidence interval, it is frequently preferred to use the alternate estimator

$$\tilde{p} = \frac{y+2}{n+4}$$

Heeding the discussion on p.208 in Samuels & Witmer (2003), we call $\tilde{p}$ the Wilson estimator of π; reasons for preferring $\tilde{p}$ to p include it's superior performance in simulation studies. As such, the 95% confidence interval is calculated via the formula $\tilde{p} \pm 1.96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$, the value 1.96 being the quantile from the standard normal distribution (z) such that the area under the normal probability distribution function between -1.96 and +1.96 is indeed 95%. Since the area under normal curve below 1.96 is therefore 0.975, we write $z_{0.975} = 1.96$. Note that this quantile can also be obtained using the R software package using the R command **`qnorm(0.975)`**.

1.4 Statistical Methods for Two or More Independent Samples

Extensions of the above one-sample methods to two independently sampled groups for continuous random variables are straightforward both in terms of confidence intervals and hypothesis testing. Interestingly, these methods apply both when we sample from two independent groups such as from male and female students at a given university, and when the two groups are created by randomization such as when volunteers are randomized to receive either Drug A or Drug B. We can better appreciate the term 'independent' in this context: in the first setting, the selection of the n_1 males in no way influences the selection of the n_2 females; in the second setting, the $n = n_1 + n_2$ randomly chosen volunteers are assigned to either Drug A or Drug B in an independent manner, such as can be thought of as the outcome of a coin toss.

With μ_1 and μ_2 representing the population means for the two groups, the relevant hypothesis test is whether the population means are equal (viz, H_0: $\mu_1 = \mu_2$, or equivalently, $\mu_1 - \mu_2 = 0$). This latter manner of writing this null hypothesis underscores that a related conclusion can be made using a confidence interval for $\mu_1 - \mu_2$, and noting whether the derived interval contains the value zero. For the hypothesis test, provided one samples from normal populations or if the two samples sizes are sufficiently large, the test statistic can be written as

$$t = \frac{\bar{y}_1 - \bar{y}_2}{SE_{\bar{y}_1 - \bar{y}_2}}$$

In this expression, $\bar{y}_1$ and s_1 are the sample mean and standard deviation for the first group, whereas $\bar{y}_2$ and s_2 are those for the second group. Further, if we assume the population variances are equal (i.e., ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$ say), then $SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{{s_P}^2}{n_1} + \frac{{s_P}^2}{n_2}} = s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$; if we assume that the population variance are unequal, then $SE_{\bar{y}_1 - \bar{y}_2} = \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$. In the former (equal-variance) case, ${s_P}^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2}$ is the so-called pooled estimator of σ^2, and the t statistic has $(n_1 + n_2 - 2)$ degrees of freedom; in the latter (unequal-variance) case, the exact number of degrees of freedom is not exactly known, and most software packages use some form of approximation such as Welch's approximation for the degrees of freedom. Additional details about this approximation (and the related Behrens-Fisher problem) may be found in the References and elsewhere. Further, an interesting connection between the equal-variance two-sample t-test and simple linear regression techniques is given in §2.5.

In the sense of robustness of results, it is probably preferred to place no assumptions regarding equality of the variances, but we hasten to add that analysis-of-variance (ANOVA) methods discussed below in §1.7 are predicated upon the requirement of equal variances. To illustrate one-way ANOVA, suppose that a random sample of n patients are randomized to receive one of three anti-hypertensive drugs (A, B, and C), such that n_1 patients receive drug A, n_2 patients receive drug B, and n_3 patients receive drug C (thus $n = n_1 + n_2 + n_3$); the balanced setting of $n_1 = n_2 = n_3 = n/3$ is not required. The relevant null hypothesis to be tested is whether the three treatment means are equal (i.e., H_0: $\mu_1 = \mu_2 = \mu_3$). In §1.7 below we discuss the relevant (F) test statistic, and in §3.6.2 we provide an extension to the unequal variance case using (approximate) likelihood methods.

1.5 Statistical Methods for Categorical (Count) Data

Categorical count data are usually – although not always – analyzed using so-called chi-square methods. Further, it can be shown that the square of a standard normal random variable has a chi-square distribution with one degree of freedom (denoted ${\chi_1}^2$), so it is not surprising that 2-by-2 (denoted 2×2) contingency tables can be analyzed using both chi-square and normal methods. It's also important to distinguish between two topics considered in introductory statistical methods: the analysis of count data given in a contingency table and goodness of fit tests (denoted GOF); both settings use chi-square test statistics, but important differences exist.

We use goodness of fit tests, for example, to test a genetic theory that progeny occur in the anticipated 12:3:1 ratio anticipated from the genetic theory and underlying Punnett square. For GOF tests, the null hypothesis is specified by the underlying theory (or by common sense), and the alternative is looking for departures from this theory or something out of the ordinary. In the 12:3:1 genetic theory case, the relevant null hypothesis is then H_0: $\pi_1 = \frac{12}{16}$, $\pi_2 = \frac{3}{16}$, and $\pi_3 = \frac{1}{16}$. Since this hypothesis gives specified values for several parameters, it is called a 'compound' hypothesis. Should we be interested in testing whether an ordinary die is fair, the null hypothesis is H_0: $\pi_1 = \pi_2 = \cdots \pi_6 = \frac{1}{6}$. The number of degrees of freedom (df) for the derived χ^2 test is one less than the number of categories, so for the genetics illustration, df = 2, and for the die example, df = 5.

Contingency tables, on the other hand, have two dimensions (and sometimes more) – i.e., one for the r rows and one for the k columns – and the cells contain count data. The simplest case is a 2x2 contingency table, so that both the row and the column variables have two levels. In cause-and-effect situations, following the methods presented in Agresti (2007), we think of the row factor as influencing

or causing the column factor; as such, this setting is analogous to linear regression discussed below in the next section. Note, however, that contingency tables and chi-square tests can also be used to simply detect an association between two variables (analogous to detecting linear correlation). Regardless of the setting, df = (r-1)×(k-1), so for a 2x2 contingency table, the associated chi-square test statistic has one degree of freedom (denoted χ_1^2).

As an example of a 2x2 contingency table, consider the following table corresponding to whether mothers smoked during pregnancy and the child's birth-weight.

		Birth Weight		**Total**
		Low	**Normal**	
Smoking Status during Pregnancy	**Smoker**	**237**	**3489**	**3726**
	Non-smoker	**197**	**5870**	**6067**
Total		**434**	**9359**	**9793**

Note that this table is of the (general) form

		Factor Y		Total
		Level Y1	Level Y2	
Factor X	Level X1	$\mathbf{n_{11}}$	$\mathbf{n_{12}}$	$\mathbf{n_{1*}}$
	Level X2	$\mathbf{n_{21}}$	$\mathbf{n_{22}}$	$\mathbf{n_{2*}}$
Total		$\mathbf{n_{*2}}$	$\mathbf{n_{*1}}$	**n**

The usual chi-square test assesses for an association between factors X and Y. As for goodness-of-fit tests, the test statistic has the form $\chi^2 = \sum \frac{(O_k - E_k)^2}{E_k}$, with O_k and E_k representing the observed and expected cell counts, and the summation is over each of the cells (for contingency tables) or categories (for goodness of fit tests).

For the smoking data, the following Minitab is obtained.

Chi-Square Test: LowBW, NormBW

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

              LowBW   NormBW  Total
Smokers         237     3489   3726
              165.13  3560.87
              31.284    1.451

Nonsmokers      197     5870   6067
              268.87  5798.13
              19.213    0.891

Total           434     9359   9793

Chi-Sq = 52.838, DF = 1, P-Value = 0.000
```

In the output, for each of the four cells, the output provides the observed counts (top), the expected counts (middle), and the contribution to the chi-square test statistic (bottom). The reported chi-square statistic, χ_1^2 = 52.838, is highly significant, so we conclude there is an association between smoking and low birth-weight.

The above results notwithstanding, we are often interested in more than simply noting such an association. In the context of this example, let π_1 denote the probability of a low birth-weight baby for the smokers, and π_2 denote the probability of a low birth-weight baby for the non-smokers. Then sometimes we want to set a CI for $(\pi_1 - \pi_2)$. The methodology to do so is straightforward and readers are referred to an introductory text such as on p.439 of Samuels and Witmer (2003).

Here, we instead consider the relative risk, $\rho = \frac{\pi_1}{\pi_2}$ and the odds ratio, $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$. Note that $\frac{\pi}{1-\pi}$ is the odds of an event. For the above smoking data, the relative risk estimate is $\hat{\rho} = \frac{0.063607}{0.032471}$ = 1.9589 since $p_1 = \frac{237}{3726}$ = 0.063607 and $p_2 = \frac{197}{6067}$ = 0.032471. Also, the estimate of the odds ratio is $\hat{\theta} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}} = \frac{237 \times 5870}{197 \times 3489}$ = 2.0240. The relative risk and odds ratio will both be nearly equal – as is the case here – when both probabilities are either very small or very large; this follows since $\theta = \rho \times \frac{1-\pi_2}{1-\pi_1}$. To interpret the estimated relative risk in the context of this setting, we say that these data suggest that the risk of having a low birth-weight baby for smokers is 1.9589 (approximately 2) times the risk of having a low birth-weight baby for nonsmokers.

Although the above methods provide us with point estimates for the relative risks and odds ratios, we are usually more interested in confidence intervals since this interval could then be used to test for a significant difference between the two groups. One method for doing so is based on a modified Wald approach; details and examples are given in Agresti (2007), and are illustrated here for the odds ratio. For the odds ratio, the relevant scale is the natural log scale, and in what follows, "log" stands for the natural (i.e., base e) logarithm; on some calculators, this is denoted "ln", short for "logarithm neperian". To obtain the 95% confidence interval for θ for the above data, we follow these steps:

1. $\log(\hat{\theta})$ = log(2.0240) = 0.7051 (again note that we use log-base-e not log-base-10)
2. $SE_{\log(\hat{\theta})} = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = \sqrt{\frac{1}{237} + \frac{1}{197} + \frac{1}{3489} + \frac{1}{5870}} = 0.0988$
3. the modified Wald 95% confidence interval for $\log(\theta)$ is 0.7051 ± 1.96*0.0988 = (0.5115 , 0.8987)
4. the modified Wald 95% confidence interval for θ is thus $(e^{0.5115}, e^{0.8987})$ = (1.67 , 2.46).

Notice that this interval does not include the value one, thereby implying inequality of the odds for the two groups; also, the entire interval lies above one implying an increase in odds for the smoking group. Further, to interpret this confidence interval: with 95% confidence, these data suggest that the odds of having a low birth-weight child for the smoking mothers is at least 1.67 times and at most 2.46 times the odds of having a low birth-weight child for the non-smoking mothers.

Note that the chi-square methods given above are only valid for large samples, that is, 'asymptotically'. For small samples, Fisher's Exact Test (FET) can be used. We suggest that the FET be used in place of the χ^2 test whenever at least 20% of the expected cell frequencies are below 5. For a 2x2 contingency table, this means that the FET should be used whenever at least one of the expected cell frequencies is less than 5. To illustrate, suppose that in the above example, it was desired to show that the risk of having a low birth-weight child was higher in the smoking group but that the data instead was as in the following (modified) table.

	Low BW	Normal	Total
Smoker	**6**	**34**	**40**
Non-smoker	**1**	**59**	**60**
Total	**7**	**93**	**100**

In this case, the following Minitab output identifies the problem – that two of the cells have expected cell frequencies below 5 – and thus that the chi-square results are invalid. It is important to point out that the observed value of 6 (and that 6 > 5) is not relevant – it is the expected cell count of $\frac{7\times 40}{100}$ = 2.8 (and 2.8 < 5) which is important in the assessment of the appropriateness of the χ^2 test.

Chi-Square Test: Low_BW, Norm_BW

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

       Low_BW  Norm_BW  Total
1           6       34     40
         2.80    37.20
        3.657    0.275

2           1       59     60
         4.20    55.80
        2.438    0.184

Total       7       93    100

Chi-Sq = 6.554, DF = 1, P-Value = 0.010
2 cells with expected counts less than 5.
```

The Fisher Exact Test fixes the row and column totals and uses the hypergeometric distribution. Calculating the p-value for the above table, one obtains p = $\frac{\binom{40}{6}\times\binom{60}{1}}{\binom{100}{7}}$ = 0.0144; this result is obtained using the following command in R: **`(choose(40,6)*choose(60,1))/choose(100,7)`**. But in calculating the p-value for this one-sided alternative, note that with the row and column totals fixed, the following table is 'more extreme' (i.e., in the direction of the alternative) that the original table:

	Low BW	Normal	Total
Smoker	**7**	**33**	**40**
Non-smoker	**0**	**60**	**60**
Total	**7**	**93**	**100**

Since no other tables can be constructed as being "more extreme than" the original table, the p-value associated with the test is the sum of the p-values associated with these two tables; thus using R:

p = **`(choose(40,6)*choose(60,1))/choose(100,7)+`**
`(choose(40,7)*choose(60,0))/choose(100,7)` = 0.0156.

At the 5% significance level, since $p < \alpha$ our conclusion is that these data suggest that a woman's smoking during pregnancy is associated with a higher incidence of having a low birth-weight baby.

To illustrate chi-square methods with a table larger than the above 2x2 contingency tables, consider the blood group – gastric ulcer data given in Aird *et al* (1954) reproduced below.

	Blood Group:				
Ulcer Status:	**A**	**B**	**AB**	**O**	**Total**
Control	**4219**	**890**	**313**	**4578**	**10,000**
Gastric Ulcer	**79**	**16**	**4**	**127**	**226**

Here, ABO blood group and gastric ulcer status were measured on 10,226 women in London to see if there was an association between ABO blood group and gastric ulcer status; the following Minitab output indicates that these data suggest that this is indeed the case ($\chi_3^2 = 10.107$, $p = 0.018$).

Chi-Square Test: Control, Gastric_ulcer

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

        Control  Gastric_ulcer  Total
A          4219             79   4298
        4203.01          94.99
          0.061          2.691

B           890             16    906
         885.98          20.02
          0.018          0.808

AB          313              4    317
         309.99           7.01
          0.029          1.290

O          4578            127   4705
        4601.02         103.98
          0.115          5.095

Total     10000            226  10226

Chi-Sq = 10.107, DF = 3, P-Value = 0.018
```

Beyond the significance test, the above Minitab output also indicates that the cells which contribute significantly to the test statistic are for the gastric ulcer group and the A and O blood types: the gastric ulcer group has significantly less individuals than expected in the A blood group and significantly more than expected in the O blood group. The term "expected" is used here to mean expected if the blood group and gastric ulcer status variables were indeed independent. This numerical result leads us to wonder if there is a genetic (or otherwise) link between the being in the O blood group and one's proneness to gastric ulcers. Interestingly, when examining the full dataset, a similar result was observed for peptic ulcers, but not for duodenal ulcers.

Note also that whenever the contingency table row and/or column variable is ordinal (instead of nominal), statistical techniques such as Mantel-Haenszel χ^2 test (for linearity) or one of the multinomial logit models (such as the proportional odds model) discussed in §4.4 should be used for testing. In this ordinal case, the usual chi-square test lacks power, and so should be avoided. In the previous example, since both variables (ABO blood group and gastric ulcer status) are nominal – and since the sample size is "large" – the data are well analyzed using the usual χ^2 test.

One important extensions of the methods considered here is to multi-way tables such as in the following 2×2×2 illustration taken from p. 209 of Agresti (2007). The following three-way tables gives data from a 1992 survey of Dayton, Ohio high school seniors who were asked whether they have ever used cigarettes (C), alcohol (A), and/or marijuana (M). The task is to detect any relationships between the three variables (A, C and M). It is sometimes advantageous to think of this 3-way table as two 2×2 "faces": the one at left corresponding to the smokers (C = Yes) and the one at right corresponding to the non-smokers (C = No).

C = Yes			C = No		
	M = Yes	M = No		M = Yes	M = No
A = Yes	911	538	A = Yes	44	456
A = No	3	43	A = No	2	279

Note that for this dataset, the estimated odds ratios for the smokers and non-smokers are $\hat{\theta}_{C=Yes} = \frac{911\times43}{3\times538} = 24.27$ and $\hat{\theta}_{C=No} = \frac{44\times279}{2\times456} = 13.46$ respectively. Even though these sample values differ, the Breslow-day test, performed in the following SAS/FREQ analysis, retains the claim of equal odds across the smoking and non-smoking faces (i.e., H_0: $\theta_{C=Yes} = \theta_{C=No} = \theta$, χ_1^2 = 0.3970, p = 0.5286). Next, note that the so-called Mantel-Haenszel (MH) estimate of the adjusted odds ratio is $\hat{\theta}_{MH} = 18.6537$ and that the provided 95% confidence interval, (7.4617, 46.6324), does not contain the value one. Similarly, the Cochran-Mantel-Haenszel conditional test of independence between A and M controlling for C (i.e., H_0: $\theta = 1$) is rejected here (χ_1^2 = 76.4645, p < 0.0001). The interpretation of this estimated odds ratio is that after controlling for whether a HS Senior is a smoker or not, the odds that he or she has used marijuana is 18.65 times higher if he or she has used alcohol at least once.

```
data one;
  do alc='yes','no';
  do cig='yes','no';
  do mar='yes','no';
    input count @@;
    output;
  end; end; end;
datalines;
911 538 44 456
  3  43  2 279
;
proc freq order=data;
  weight count;
  tables cig*alc*mar/
    nopercent nocol
    norow cmh;
run;
```

```
                    The FREQ Procedure
            Summary Statistics for alc by mar
                  Controlling for cig

   Cochran-Mantel-Haenszel Statistics (Based on Table Scores)
 Statistic    Alternative Hypothesis    DF       Value      Prob
 ---------------------------------------------------------------
     1        Nonzero Correlation        1     76.4645    <.0001
     2        Row Mean Scores Differ     1     76.4645    <.0001
     3        General Association        1     76.4645    <.0001

        Estimates of the Common Relative Risk (Row1/Row2)
Type of Study   Method                 Value     95% Confidence Limits
---------------------------------------------------------------------
Case-Control    Mantel-Haenszel      18.6537      7.4617      46.6324
(Odds Ratio)    Logit                19.1174      7.7212      47.3339
Cohort          Mantel-Haenszel      10.4729      4.4026      24.9132
(Col1 Risk)     Logit                10.5865      4.4592      25.1332
Cohort          Mantel-Haenszel       0.8199      0.7971       0.8434
```

```
(Col2 Risk)     Logit                 0.8624      0.8387      0.8867

        Breslow-Day Test for
    Homogeneity of the Odds Ratios
    ------------------------------
    Chi-Square              0.3970
    DF                           1
    Pr > ChiSq              0.5286
        Total Sample Size = 2276
```

A few comments are in order here. First, for more complicated datasets and analyses, we will use the SAS package, giving both the program (left) and output (right) so as to familiarize readers with the syntax and output. Second, we stress that the above Mantel-Haenszel odds ratio estimate is indeed a "conditional" one in that the relationship between alcohol and marijuana usage have been assessed after adjusting (or controlling) for smoking status. As such, $\hat{\theta}_{MH}$ is a weighted average of $\hat{\theta}_{C=Yes}$ and $\hat{\theta}_{C=No}$. A less-desirable result would be obtained by summing together (i.e., "collapsing") the two faces to obtain the following "marginal" table and marginal odds ratio estimate: $\hat{\theta}_{MARG} = \frac{955\times322}{5\times994} = 61.87$.

Marginal Table	M = Yes	M = No
A = Yes	955	994
A = No	5	322

Instead of adjusting for the third variable (smoking status), the above marginal result simply ignores it, which can often give unreliable results.

1.6 Simple Linear Regression Methods

Two continuous random variables – one called the dependent variable (Y) and the other the independent variable (X) – can be related using simple linear regression (SLR). SLR holds that

$$E(Y_k) = \beta_0 + \beta_1 x_k \text{ , for } k = 1, 2, \ldots, n$$

This model makes several underlying assumptions (including normality and constant variance). In this case, ordinary least-squares and maximum likelihood estimation methods are equivalent and can be used to estimate the unknown intercept (β_0) and slope (β_1). Testing whether the slope is zero ($\beta_1 = 0$) is equivalent to testing that the independent variable is not a good predictor of the dependent variable. A related measure is the linear correlation coefficient (denoted ρ); since it can be shown that $\beta_1 = \rho \times \frac{\sigma_Y}{\sigma_X}$, testing whether the correlation coefficient is zero is equivalent to testing whether the slope is zero. Setting a confidence interval for the correlation coefficient is more involved than doing so the slope parameter, and we addressed these topics in depth in the next chapter.

Interestingly, as discussed on p.577 of Samuels & Witmer (2003), regressing Y on a dummy variable is equivalent to the equal-variance, two-sample t-test discussed above in §2.4. To see this, consider testing for equality of two means associated with drugs A and B in a study in which n_1 patients are randomized to receive drug A and n_2 patients are randomized to receive drug B; the null hypothesis is then H_0: $\mu_A = \mu_B$. Define the so-called dummy variable (D_k) for each patient to be one for the drug A patients and zero for the drug B patients. Thus,

$$D_k = \begin{cases} 1, \text{for } k = 1, 2, \ldots n_1 \\ 0, \text{for } k = n_1 + 1, n_1 + 2, \ldots\ n_1 + n_2 \end{cases}$$

Then for the SLR model, $E(Y_k) = \beta_0 + \beta_1 D_k$, for k = 1, 2, …, $n = n_1 + n_2$, the right hand side of this expression is equal to $(\beta_0 + \beta_1)$ for the n_1 patients who received Drug A and is equal to β_0 for the n_2 patients who received Drug B. The left hand sides in these situations are equal to μ_A and μ_B respectively. Therefore testing whether the slope parameter (β_1) is equal to zero is equivalent to testing H_0: $\mu_A = \mu_B$ since $\beta_1 = \mu_A - \mu_B$. In sum, simple linear regression with a dummy variable in place of the independent variable is equivalent to the equal-variance two-sample t-test. An illustration is given in §2.5. Not surprisingly, one-way ANOVA with d drugs extends this idea by introducing (d-1) dummy variables (in place of the single dummy variable introduced above) and using multiple linear regression in place of SLR.

1.7 One- and Two-way ANOVA

The final introductory topic considered here is one- and two-way ANOVA (analysis of variance). An illustration of one-way ANOVA would involve comparing the means efficacy (as measured using some quantitative measurement) of three the drugs: A, B and C. The null hypothesis states that these three means are equal. This hypothesis is tested using the usual F statistic, which is obtained by forming the ratio of the between groups mean squares over the within groups mean squares. When this F statistic is sufficiently large, we conclude that the (between groups) drug means significantly differ. A next step is to then perform a mean separation procedure. These include Tukey's HSD test, Fisher's LSD test, the Newman-Keuls test, or to use Bonferroni procedure. Each of these methods helps to decide precisely which means differ from one another, and additional details can be found in most introductory textbooks.

A two-way ANOVA analysis would be appropriate for example if we wanted to study the same three drugs (i.e., drugs A, B, and C) and two methods of delivery (tablet versus injection). Here, "drug" and "method of delivery" are called **factors**, whereas A, B, C, tablet, and injection are called **levels** of these factors. A big concern in studies of this type is whether interaction exists between the factors drug and method of delivery (as would occur, for example, if drug B is "best" when taken in tablet form but that no differences exist between the drugs when delivered via injection). We return to these concepts in more detail in §3.2.

1.8 References

Agresti, A. 2007, *An Introduction to Categorical Data Analysis*, 2nd Edition, Wiley.

Aird, I., Bentall, H.H., Mehigan, J.A. & Roberts, J.A.F., 1954, "The Blood Groups in Relation to Peptic Ulceration and Carcinoma of Colon, Rectum, Breast, and Bronchus: An Association between the ABO Groups and Peptic Ulceration," *British Medical Journal*, 2(4883), 315-321.

Samuels, M.L. & Witmer, J.A., 2003, *Statistics for the Life Sciences*, 3rd Edition, Pearson/Prentice-Hall.

Wackerly, D., Mendenhall, W. & Scheaffer, R.L., 2008, *Mathematical Statistics with Applications*, 7th Edition, Brooks/Cole/Duxbury.