Chapter 1. Quick Review of Basic Biostatistics – ***students are strongly advised to work though an introductory text as well!***

1. (Random) Variables (RVs) are
   - quantitative
     - discrete
     - continuous
   - qualitative
     - nominal
     - ordinal

2. Examples of RVs are number of head's obtained when you toss a coin 100 times (discrete), your exact age (continuous, if you don't round off), your hair color (nominal), your understanding of molecular genetics if you choose from the list: excellent, good, moderate, fair, poor, horrible, non-existent.

3. Quantitative RVs have distributions. Examples of discrete distributions are the binomial, Poisson, geometric (number of coin tosses necessary to get the first head), uniform (on the integers 1, 2, …, n). Examples of continuous distributions are the normal (or Gaussian), chi-square, t-, F-, a continuous version of the uniform, etc. These distributions have probability functions (called pdf's for probability density functions) associated with them; for the Binomial distribution with n independent trials and with success probability π, for example, the pdf is

$$P(y) = {}_nC_y \, \pi^y (1-\pi)^{n-y}, \text{ for } y = 0, 1, 2, \ldots, n.$$

   Distributions also have so-called moments associated with them (in some cases some of these are infinite though); the first moment is the mean (μ) and the second (centralized) moment is the variance (σ^2). For the Binomial, the $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$. For other distributions, these formulas don't hold, but their moments can still be obtained using the definitions $\mu = E(Y)$ and $\sigma^2 = E[(X - \mu)^2]$.

4. If we sample (randomly) from a known distribution, we are often interested in understanding what to expect for the distribution of a statistic such as the sample proportion (denoted "p" here) or a sample mean (denoted "y-bar" here). The Central Limit Theorem (CLT) tells us that as long as the sample size is large enough, these distributions will be approximately bell-shaped (normal) even when the underlying population distribution is not. The CLT goes on to tell us about the mean and standard error (counterpart of the SD) of these distributions; Samuels & Witmer cover this in Chap. 5.

5. The above scenario is unrealistic since we never (rarely) know the underlying population distribution, but it is helpful since, by reversing the inherent logic, we can set confidence intervals (CI's) or perform hypothesis tests (HT's) regarding the population parameters based on what we see in our sample. For example, in one (random) sample where we record a quantitative rv, we can set a CI for the true mean (μ) by using the usual t-distribution; similarly we can test (HT) whether this unknown population mean is equal to a specific number again using the t- test statistic. 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" 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. Sometimes, for a single sample involving the Binomial setup, we are interested in setting a CI for the unknown success probability, π. Instead of the usual estimator, p = y/n, students are usually perplexed as to why the CI is based on another estimator, p-tilde = (y+2)/(n+4). We return to this situation and give the motivation for why the latter (p-tilde) is better than the former (p) in Chapters 2 and 3.

6. An easy extension of the above (paragraph 5) occurs when we want to compare (via a CI or HT) the (population) means of two independent groups (e.g., M's versus F's, those who got drug A versus those who got drug B, etc.). In this case, our CI or HT is based on the two independent sample t-statistic, with two variants: one where we can assume the population variances are the same (less common, I suppose) and one where we cannot. Extensions of the latter situation (unequal variance two-independent sample t-test) does not generalize as easily to one-way ANOVA and simple linear regression as the former situation, although we will manage to do so using likelihood-based asymptotic tests and methods.

7. Categorical count data are usually (though not always) analyzed using chi-square methods. It is important to remember that the square of a standard normal RV has a chi-square distribution with one df (denoted χ_1^2), so it is not surprising that 2x2 contingency tables can be analyzed using chi-square and normal methods. It's important to distinguish between the analysis of contingency tables of count data and goodness of fit (GOF) tests since both are based on chi-square statistics. We use GOF tests, for example, to test a genetic theory that progeny occur in the anticipated 12:3:1 ratio anticipated from our genetic theory and the underlying Punnett square. For GOF tests, the null hypothesis is specified by the underlying theory (or common sense), and the alternative is looking for departures from this theory. Contingency tables, on the other hand, have two dimensions (sometimes more!): for rows and columns, and the cells contain count data – the easiest case is a 2x2 CT. As an example, 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**

Our usual chi-square test tests for an association for factors X and Y. For the smoking data, the following Minitab is relevant.

Chi-Square Test: LowBW, NormBW

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

         LowBW   NormBW  Total
    1      237     3489   3726
        165.13  3560.87
        31.284    1.451

    2      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
```

The chi-square statistic here (52.838) is highly significant, so we conclude there is an association between smoking and low birth-weight, but we are usually interested in more than this. If we define π_1 as the probability of a low birth-weight baby for the smokers, and π_2 as the probability of a low birth-weight baby for the non-smokers, then sometimes we want to set a CI for $(\pi_1 - \pi_2)$. Since this procedure is discussed at length in introductory texts (see Samuels and Witmer, p.439), consider instead the relative risk, $RR = \pi_1 / \pi_2$, and the odds ratio, $OR = [\pi_1 / (1 - \pi_1] / [\pi_2 / (1 - \pi_2]$; sometimes OR is expressed using θ. For the above date, the estimate of RR is 0.063607 / 0.032471 = 1.9589 (approximately 2) since $p_1 = 237/3726 = 0.063607$ and $p_2 = 197/6067 = 0.032471$. For the same data, the estimate of the odds ratio (θ) is $q = (n_{11} * n_{22})/(n_{12} * n_{21}) = (237*5870)/(197*3489) = 2.0240$ (also approximately 2). We set a 95% CI for θ using the following procedure: $\log(2.0240) = 0.705096$, and $SQRT\{1/237 + 1/197 + 1/3489 + 1/5870\} = 0.098755$. Thus, a 95% CI for $\log(\theta)$ is

$$0.705096 +/- 1.96*0.098755 \quad \text{or} \quad (0.511536 , 0.898656).$$

Finally, the 95% CI for θ is $(\exp\{0.511536\} , \exp\{0.898656\}) =$ **(1.668 , 2.456)**. Note that this interval does not include unity (1).

8. The chi-square test in paragraph (7) is only valid asymptotically (for large samples), and Fisher's Exact Test (FET), discussed in Samuels and Witmer p.432ff, is used for small samples or in situations where a nonparametric test is desired. This test is based on the hypergeometric distribution, and is useful for drug testing involving human subjects.

9. 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), which posits

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

This model makes several underlying assumptions (of which students should be aware!), and ordinary least-squares (OLS) or equivalently maximum likelihood (ML) methods are 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 concept is the linear correlation coefficient, ρ, and we note that $\beta_1 = \rho * (\sigma_Y / \sigma_X)$, so we can see that testing whether the slope is zero is equivalent to testing whether the correlation coefficient is zero. Unfortunately, more involved methods are needed for setting a CI for ρ, and these are addressed in the next Chapter.

10. Interestingly, regression on a dummy variable is the equal variance independent sample t-test (see Samuels and Witmer, p. 577). More specifically, consider the following setup involving testing for equality of two means associated with drugs A & B, where n_1 patients are randomized to receive drug A, and n_2 patients are randomized to receive drug B. Here, the relevant null hypothesis is H_0: $\mu_A = \mu_B$. Next, define the ***dummy*** variable,

$$D_k = 1 \text{ for } k = 1, 2, \ldots, n_1 \text{ (drug A)}$$
$$= 0 \text{ otherwise } (n_2 \text{ patients; drug B})$$

Then for the SLR model

$$E(Y_k) = \beta_0 + \beta_1 D_k \text{ , for } k = 1, 2, \ldots, n = n_1 + n_2,$$

the right hand side 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. Further, the left hand side is 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 $\beta_1 = (\mu_A - \mu_B)$. Thus, SLR with a dummy variable is equivalent to the equal-variance independent sample t-test. One-way ANOVA with d drugs extends this idea by introducing (d-1) dummy variables.

11. The last topic covered in introductory biostatistics courses is one- and two-way ANOVA. An illustration of one-way ANOVA involves comparing the means (e.g., of efficacy as measured using some quantitative measurement) of three drugs, A, B and C. The null hypothesis states that these three means are equal, this hypothesis is tested using the usual F statistic (for Fisher), and when it is rejected, mean separation procedures (MSP's – also called multiple comparison procedures, MCP's) such as Tukey's Honestly Significance Procedure, Fisher's Least Significance Difference procedure, the [Student] Newman-Keuls method, the Bonferroni procedure, or any of a long host of methods to decide which means differ from one another; the MCP's are discussed in Samuels and Witmer, section 11.8. A two-way ANOVA analysis would be appropriate if we wanted to study three drugs (A, B, C) and two methods of delivery (tablet versus injection). Here, "drug" and "method of deliver" are called ***factors***, and A, B, C, tablet, and injection are called ***levels*** of these factors. We take up these concepts in more detail in Chapter 3, but interested students can check out Samuels and Witmer's discussion of two-way ANOVA and especially interaction in their section 11.7.