Solutions to Biostatistics and Epidemiology Step 1 Sample Questions Set 1

1. a. The correct answer is **A**. Birthweight measured in grams is a variable measured on a continuum. This is a continuous variable.
   b. The correct answer is **B**. Birthweight classified as low, medium and high is a classification that has a natural ordering. Hence this is an ordinal variable.
   c. The correct answer is **D**. Birthweight classified as low versus not low is dichotomous or binary. A dichotomous variable is one which measures the presence or absence of a trait.
   d. The correct answer is **C**. Delivery type measured as cesarean, natural, or induced is a categorical variable, but the levels of the classes have no natural ordering. Therefore, this is a nominal categorical variable.
2. The correct answer is **A**. The average number of lesions can be calculated directly as (0 x 50 + 1 x 30 + 2 x 10 + 3 x 10)/100 = 0.8.
3. The correct answer is **B**. Because the median is different from the mean, we know the data are skewed. In fact, since the mean (67) is smaller than the median (76), we know that the data are negatively skewed – the tail extends to the left. Remember that the mean is influenced by outlying observations and is pulled in the direction of the tail. The only way for the mean to be smaller than the median is if the data are negatively skewed. When data are skewed, the correct measure of central tendency is the median.
4. The correct answer is **C**. We know that when data are approximately normally distributed, 68% of the data fall within ±1SD of the mean, 95% of the data fall within ±2SD of the mean, and >99% of the data fall within ±3SD of the mean. Therefore the correct answer is determined by 230 – 2 x 10 and 230 + 2 x 10, or 210 and 250.
5. The correct answer is **D**. A correlation coefficient is a measure of the strength of the linear association between two continuous variables. The closer the correlation coefficient is to +1, the tighter the scatter of points around a positively sloped line. The closer the correlation coefficient is to -1, the tighter the scatter of points around a negatively sloped line.
6. The correct answer is **B**. A case-control study would be the best study design in this case because the disease prevalence is 1 in 1000. To conduct a cohort study, we would need to enroll thousands of patients to observe even a small number of cases. The cohort study is therefore very costly and inefficient.
7. The correct answer is **C**. In a case-control study, it is important to select the control group so they are as similar as possible to the cases in every respect except for their disease status. By selecting the controls from the same hospital's outpatient clinic, we are selecting controls from a population of patients that are likely to similar to the cases. The hospital serves as a surrogate of sorts to facilitate selection of a control subjects that are more likely to be similar with respect to socio-demographic and other exposure factors.
8. The correct answer is **B**. This is a cohort study. A cohort study is characterized by selection of subjects based on exposure status (exposed or not exposed) and then following those subjects forward in time to determine disease status. These characteristics are present in this study design.
9. The correct answer is **A**. This is an example of a classic case-control study design. Subjects are selected based on their disease status and their exposure status is then ascertained. An additional clue that this is a case-control study is the fact that an odds ratio of exposure among the cases compared to exposure among the controls is the measure of association that is calculated.
10. The correct answer is **C**. This is a longitudinal study, meaning that we are simply following a group of people forward in time (note that a cohort study <u>is</u> an example of a longitudinal study). However, this is not a prospective cohort study because this study lacks an unexposed group – every cohort study must have both an exposed and an unexposed group. Therefore, this is simply a longitudinal study.
11. The correct answer is **E**. This is a double-blind randomized trial. The double-blind means that neither the investigators nor the patients know what drug they receive. The fact that the study is randomized means that patients are assigned to the treatment using a process that is random rather than systematic - for example, if the investigator got to choose the treatment for each subject.
12. The correct answer is **E**. Prevalence is the total number of cases divided by the total population size. In this problem there are 10,000 cases in a population of size 250,000, so the prevalence is given by choice E.
13. The correct answer is **A**. 'New cases' refers to incidence, which is a measure of the rate of new cases of disease.
14. The correct answer is **E**. Prevalence is the total number of cases divided by the total population size. The total number of cases is captured by all the cases in the first column of the table, W + X. The total population size is the sum of all the cell counts, W + X + Y + Z. Therefore, the prevalence is given by choice E.

15. The correct answer is **D**. We're interested in the rate of new cases in a single year. If there are 10,000 people but 1000 are already infected at the start of the year, then there are only 9,000 people at-risk of becoming infected in the coming year. In that year, 100 people become infected. Therefore, the annual incidence rate is 100/9,000.
16. The correct answer is **D**. The increasing rate of vaccination over time means that fewer people will become infected with hepatitis B. Therefore, incidence goes down.
17. The correct answer is **D**. You might find this one easier to answer if you set up a 2 X 2 table based on the information given.

	Acute coronary syndrome	No acute coronary syndrome	TOTAL
High fibrinogen	40	30	70
Normal fibrinogen	20	40	60
TOTAL	60	70	130

The question asks for the estimate of relative risk, which is the ratio of the disease rate among the exposed to the disease rate among the unexposed ('exposure' here means a high baseline fibrinogen level). Therefore the relative risk = (40/70)/(20/60).

18. The correct answer is **D**. Since this is a case-control study, then we cannot calculate a relative risk. For a case-control study, the only measure of association you can calculate is an odds ratio which is choice D.
19. The correct answer is **B**. The odds ratio for a case-control study is the odds of exposure among the cases divided by the odds of exposure among the controls. From the table given, the odds of being a smoker among the cancer patients is 50/40 while the odds of being a smoker among the non-cancer patients is 60/80. Therefore, the odds ratio is (50/40)/(60/80).
20. The correct answer is **A**. First, in this example notice that we have odds ratios reported for more than two groups. Here, BMI is categorized into four groups, and then the fourth group is further sub-classified into three groups. When this is the case, we need to pick a category to serve as the 'reference' group – that is, the group to which all others will be compared. In this table, the normal weight subjects serve as the reference category, and all other categories are compared to that group. How can we tell? The reference group is always identified by reporting a 1.0 for the ratio measure of association, and there is no 95% CI reported. So in this table, normal weight subjects serve as the reference group. Working our way down the table, we interpret each of the ORs as follows:
    - Underweight individuals' odds of having major depression are 17% higher than the odds of having major depression for individuals with normal weight (OR = 1.17)
    - Overweight individuals' odds of having major depression are 14% lower than the odds of having major depression for individuals with normal weight (OR = 0.86).
    - Obese individuals' odds of having major depression are 88% higher than the odds of having major depression for individuals with normal weight (OR = 1.88).
    - Class 1 obese individuals' odds of having major depression are 28% higher than the odds of having major depression for individuals with normal weight (OR = 1.28).
    - Class 2 obese individuals' odds of having major depression are 76% higher than the odds of having major depression for individuals with normal weight (OR = 1.76).
    - Class 2 obese individuals' odds of having major depression are approximately five times higher than the odds of having major depression for individuals with normal weight (OR = 4.98).
21. The correct answer is **C**. As in the previous problem, a RR = 1.0 in a table such as this indicates the reference category. That said, relative to the lowest quintile of the distribution of CRP, as CRP increases, the risk of heart attack/stroke increases as evidenced by the increasing RR values.
22. The correct answer is **B**. Bias is any systematic error which results in either an over- or under-estimate of the effect of interest. Therefore, the correct answer is B. Confounding (choice A) is a specific type of bias that results when a variable masks the true association between disease and exposure. Age is a common confounder. If an investigator is not careful in making sure the distribution of age is comparable between the arms of a clinical trial, it is possible that any affect being attributed to the drug may be due to the difference in age between the two groups. Interaction (choice C) is not a form of bias at all, but rather the situation in which the relationship between exposure and disease changes depending on the value of another variable. For instance, a targeted

cancer drug may be highly effective in patients with a given mutation but completely ineffective in patients lacking the mutation. The drug-disease association changes depending on the mutation status of the patient. This is an interaction. Stratification (choice D) is an analysis method to deal with confounding.

23. The correct answer is **C**. Double blinding is to avoid the bias that can be induced by the observer (investigator) knowing which arm the patient is randomized to, and to avoid bias on the part of the subject. You can imagine that if the investigator knows which arm of the trial the patient is randomized to, that this might influence the investigator's expectation or anticipation of a response. The investigator might look harder for an effect in patients he/she knows are receiving the investigational agent, and likewise might look less diligently for an effect in patients randomized to placebo or standard of care. The same is true of the patient. Therefore, double-blinding helps to mitigate this kind of bias.
24. The correct answer is **E**. When subjects are randomized, you are relying upon chance to achieve a balance of potentially confounding factors across the arms of the study. For example, randomization results in relatively equal distributions of age, race, gender, etc. across the two study arms. Most of the time, the randomization process is successful in achieving an equitable distribution of these factors. If an investigator is especially concerned that a particular factor be equally distributed across the two study arms, then the investigator can employ a stratified randomization procedure to enforce equal numbers of patients within certain subgroups in each of the study arms.
25. The correct answer is **B**. When patients self-report their clinical outcome in a clinical trial, it is especially important to make sure blinding is used. Otherwise, patients in the investigational agent arm might be more likely to report a positive clinical outcome, whereas patients who know they are in the placebo arm might over-report incidents of no clinical benefit. To avoid this problem for this study, the investigators should blind the patients to their randomization status.
26. The correct answer is **B**. The analysis technique described in this example is called a stratified analysis, which is an analytic technique to control for confounders. Here, the investigators are interested in the association between alcohol consumption and lung cancer, and they observe an association. However, smoking is confounder here. We know smoking is a strong risk-factor for lung cancer, and furthermore, people who consume a lot of alcohol are more likely to smoke. When we break up the population into two groups – smokers and non-smokers – and then calculate the ORs for the alcohol-lung cancer association within each of these groups – the association goes away in both groups. This is called a stratified analysis, and is a classic example of how to diagnose a confounder.
27. The correct answer is **E**. This is a classic example of observer bias. If the investigators are trying to establish an association between race and end-stage renal disease based on the pathology of the sample, then knowing the race of the subject for the sample being examined can lead an investigator to look for the presence of disease in support of the hypothesis.
28. The correct answer is **B**. This is a classic example of recall bias. Mothers of affected children are much more likely to scour the past and search for an exposure that caused the event than mothers of unaffected children.