BIOSTATISTICS 661
Things to Keep in Mind

1. Cell phones are very disruptive to the class. Before you enter the classroom, turn your cell phone *completely off*. Better yet, remove the battery. Better yet, do not bring it to class at all.

   If your cell phone rings during class, then simply carry out your belongings and leave the classroom immediately. No excuses will be accepted.

2. The argument "I followed the right procedure, but got the wrong answer" is not acceptable in this class. If you follow the right procedure you *will* get the right answer. If you got the wrong answer, then you *must* have followed the wrong procedure. There will be no partial credit in those instances. This class is about the *ability to follow the right procedures and carry them all the way through to conclusion.*

   If the final answer you get is obviously wrong, there will be zero credit, no matter what the reason is. As an example, suppose there is a parameter $\theta < 0$, and the variance of some random variable came out as 2θ. The credit is zero. Why? The quantity 2θ is negative and the variance can not be negative. The answer 2θ is *obviously wrong.* Failing to observe that something is obviously wrong more than wipes out all potential credit you may have gotten otherwise. Always verify that your final answer is not obviously wrong.

   Another example of obviously wrong answers involves obtaining function arguments outside their domains. Examples: $\sqrt{x^2-1}$ where $x^2 < 1$ is feasible. $\log x$ where $x < 0$ is feasible; $\arcsin(x^2+2)$ (with real x).

3. Most of the derivations in this class involve applying basic theorems or formulae to a specific problem. Most problems involve one or more of the following: 1) algebraic manipulation, 2) taking derivatives and integrals, 3) manipulating inequalities, 4) taking limits. Making a mistake in those derivations counts as the wrong procedure. If you are uncomfortable with such derivations, then you are not prepared to take this class.

   At a higher level, there is the mathematical logic and long-chain reasoning. This is even more important than the mechanical skills listed in the last paragraph.

4. Simple quick checks on derivations are often possible. Is the pdf or pmf ≥ 0? Does it integrate (or sum) to 1? Do function arguments make sense? e.g. $\log(x)$ when $x < 0$; $\sqrt{x}$ when $x < 0$; $\arcsin(x)$ when $|x| > 1$. Is the variance ≥ 0? Are probabilities between 0 and 1? The expected information is a variance and can not be negative.

5. Mathematics is the language of statistics. Learning statistics without learning its language is virtually impossible. It is important to have a good understanding of basic mathematical notation and how things are written so that ambiguity is eliminated or, at least, greatly reduced.

   One example is how to define or specify a *function.* We will use the following approach: we specify a set $\mathcal{X}$ and a rule that associates with each $x \in \mathcal{X}$ a value $y \in \mathcal{Y}$. The set $\mathcal{X}$, known

as the *domain of f*, must be specified explicitly. The set $\mathcal{Y}$, known as the *range of f*, is often inferred. Example: The following does not define a function

$$g(x) = x^2,$$

while the following does

$$f(x) = x^2, \qquad 0 < x < 2.$$

The domain of f is $\mathcal{X} = (0, 2)$, and it is inferred that the range is $\mathcal{Y} = (0, 4)$. Some of the functions we deal with in this class are: pdf, pmf, mgf, cdf, likelihood, log-likelihood, power, coverage probability.

Failure to specify function domains is a BIG source of mistakes in transformations of random variables, as it leads to integration over the wrong ranges and eventaully leads to the wrong answers. You have been warned!

6. A common, and lethal, mistake is ignoring the ranges of transformed random variables. If $(U, V) = g(X, Y)$, finding the joint density of (U, V) is relatively easy. However, to integrate out v, one needs to know the limits of integration which generally depend on u. The range (sample space) for (U, V) should be written as a joint range for (u, v), not separately for u and v (since the range of u may depend on v, and vice versa).

   Whenever you write down or derive a density or a pmf, write the range next to it. Make it a habit. Look at C&B carefully and you will see that the range is always given as part of the specification of pdfs and pmfs. Example:

   $$f(x) = \theta^{-1}, \qquad \theta > 0$$

   is not a pdf, while

   $$f(x) = \theta^{-1}, \qquad 0 < x < \theta, \theta > 0$$

   is a pdf.

   This is not mere semantics. Ignoring these simple rules can, and often does, lead to grave mistakes. You have been warned, twice!

7. Planning: Most problems can be solved in more than one way. It is always worth spending a little extra time upfront to choose the easiest possible route.

8. Quite often, a new problem can be transformed into or related to an old problem that we know how to solve.

9. Read problems carefully. Sometimes, the way the problem is stated is itself a hint. Hints are not always marked as "Hint:". Words like "exact" and "approximate" are important. If the problem says "exact confidence interval", it is telling you "do not use the CLT". If it says "approximate confidence interval", then you should probably not try tedious exact calculation, and consider the CLT or some other approximations.

10. For continuous random variables, the density is **NOT** a probability, and should **NOT** be interpreted as a probability. For any continuous random variable X, the probability of any single value x is **zero**, $P(X = x) = \int_x^x f_X(x)dx = 0$.

11. The method of Jacobians does **NOT** apply to discrete random variables.

12. Someone asked why don't we use a "formula" to define the order statistics. For continuous variables (no ties), the i-th order-statistic $X_{(i)}, 1 \leq i \leq n$, can be written as an explicit function of $X_1, \cdots, X_n$ as follows

$$X_{(i)} = \sum_{j=1}^{n} X_j I\left(i = \sum_{k=1}^{n} I(X_k \leq X_j)\right).$$

This is one instance in which a few words in simple English can be much more informative than a formula.

13. When working with transformations from $X = (X_1, X_2)$ to $Y = (Y_1, Y_2)$ using Jacobians: (a) If we partition, we partition the sample space of X, not Y. (b) If we partition, we partition based on both x_1 and x_2, not one of them alone, e.g.

$$A_1 = \{(x_1, x_2) : -\infty < x_1 < 0, -\infty < x_2 < \infty\}$$

NOT $A_1 = \{x_1 : -\infty < x_1 < 0\}$ even if the sets $A_1, \cdots$ are actually determined by the value of x_1 alone.

14. Notation: Is every symbol defined? If a symbol pops up in the middle of a derivation, there must be a good reason and a clear justification.

15. If X_n converges in distribution to X, it would be wrong to say that "we approximate X_n by X" since convergence in distribution says nothing about the distance $|X - X_n|$. However, we can say that we approximate the cdf of X_n by the cdf of X if n is large. The normal approximation based on the CLT is a special case of this in which the limit cdf is $\Phi(\cdot)$.

16. When considering limits of sequences, say as $n \to \infty$, "n" itself should not appear anywhere in the limit value (after the arrow). Example: Let $a_n = 1/n + e^{-n}, n = 1, 2, 3, ...$ Then

$$\begin{aligned} a_n - 1/n \to 0 \quad & \text{as } n \to \infty \quad \text{is correct,} \\ a_n \to 1/n \quad & \text{as } n \to \infty \quad \text{is wrong,} \\ a_n - e^{-n} \to 0 \quad & \text{as } n \to \infty \quad \text{is correct,} \\ a_n \to e^{-n} \quad & \text{as } n \to \infty \quad \text{is wrong.} \end{aligned}$$

This applies to other types of limits,

$$\begin{aligned} \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathrm{n}(0, \sigma^2) \quad & \text{as } n \to \infty \quad \text{is correct,} \\ \bar{X}_n - \mu \xrightarrow{d} \mathrm{n}(0, \sigma^2/n) \quad & \text{as } n \to \infty \quad \text{is wrong.} \end{aligned}$$