Statistics 403 Practice Exam 1

- During the actual exam you will not be able to use notes, books, etc.
- These are practice problems for the exam. The actual exam will not be this long. Also, I have not included practice problems that are very similar to the homework problems, so you should study those too. I will not put problems with lengthy calculations on the exam, but you should still study all the problems from problem sets 1-6 including the solutions to review the main ideas of the course so far.
- If you are asked for a probability that cannot be calculated by hand, you can express your answer either in terms of normal probabilities (e.g. $P(Z < \ldots)$, $P(Z > \ldots)$), or in terms of R commands (e.g. `pnorm(...)`).

1. Suppose our goal is to estimate the expected value μ of a population, and to provide a 95% confidence interval around our estimate. The standard deviation of the population is 2. If we aim to have a confidence interval that is around 0.25 units wide, what sample size is required?

2. Suppose we have a test statistic T that is standardized under the null hypothesis. We observe a test statistic value of $T = 2.2$. What is the two-sided p-value for our data?

3. Suppose we observe data $X_1, \ldots, X_n$, and use it to form a 95% confidence interval for the expected value of the population. Rather than using the sample standard deviation to form the interval, we use the "nominal value" $\sigma = 1.5$ that is based on previously collected data of a similar type. However the truth is that $\sigma = 2$. What is the actual coverage probability of our confidence interval?

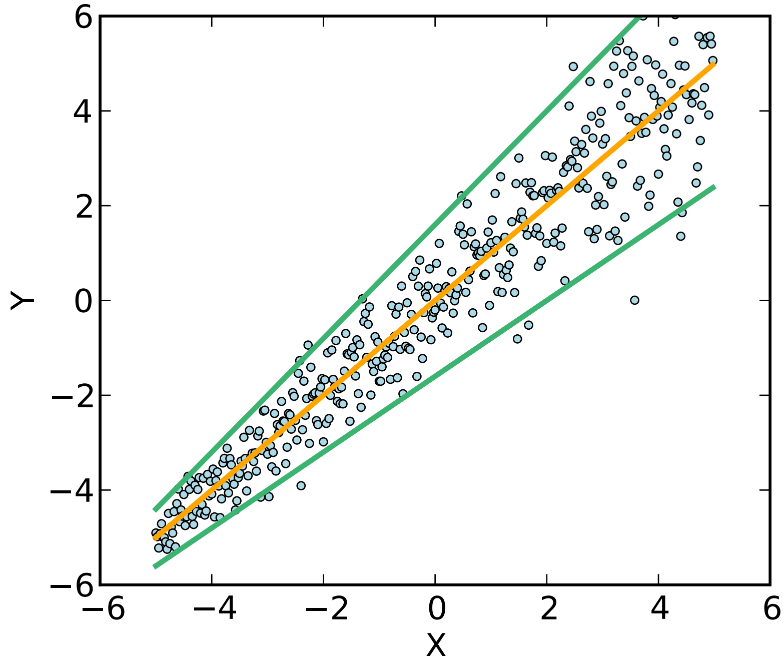
4. Suppose we observe $X_1, \ldots, X_n$ from a population with mean μ and standard deviation $1/2$. For every pair of distinct observations X_i, X_j, the correlation coefficient between the observations is $\text{cor}(X_i, X_j) = 0.4$.
   (a) What is the covariance between each pair of distinct observations?
   (b) What is the variance of the sample mean of these data?
   (c) Now suppose we are able to obtain an independent sample from a population with the same mean and standard deviation as this one. What is the variance of the sample mean in this case?

5. Suppose we are studying a quantity X that follows a normal population with mean μ and variance 1. However, we are sampling in such a way that negative values can never be included in the sample (apart from this, the sample is representative of the population). We use the sample mean to estimate the expected value as usual. Let $B(\mu)$ be the bias of our estimate. Answer the following two questions, giving a brief explanation of your reasoning.

(a) As a function of μ, is $B(\mu)$ positive, negative, or zero?

(b) As a function of μ, is $|B(\mu)|$ decreasing, increasing, or constant?

6. The scatterplot below shows a sample from the joint distribution of Y and X, which are continuously distributed values.



(a) Is the conditional mean $E(Y|X)$ an increasing, decreasing, or constant function of X?

(b) Is the conditional variance $\text{var}(Y|X)$ an increasing, decreasing, or constant function of X?

7. Suppose we are comparing two populations: the treated population X has standard deviation 1, and the control population Y has standard deviation 2. Our budget allows us to collect a total sample size of 30 units. To maximize the power, should we collect (i) the same number of treated and control samples, (ii) more treated and fewer control samples, or (iii) more control and fewer treated samples? Briefly explain your reasoning, but you do not need to derive a formula for the optimal design, or provide numerical calculations of the power.

8. Suppose we observe patients who have been given one of two treatments, A and B. The patients' responses are categorized as "no response" (NR), "partial response" (PR), and "complete response" (CR). The following table gives the distribution of treatment responses for each treatment type:

	NR	PR	CR
A	0.3	0.5	0.2
B	0.2	0.2	0.6

Suppose that in the medical center we are studying, 30% of patients are treated according to treatment A, and the remainder are treated according to treatment B.

(a) What is the marginal (overall) probability that a patient in this medical center will have no response to treatment?

(b) Among all patients having a complete response, what proportion are treated with treatment A?

(c) What is the probability that a person selected at random from this medical center will have been treated with treatment B and will have had a partial response?

9. Suppose we are studying the protein yield of soybeans. The particular population of soybeans we are studying is a mix of two strains. Strain A has mean yield 10 and standard deviation 3, and strain B has mean yield 12 and standard deviation 4. An individual soybean plant in our study has probability p of being strain A, and probability $1 - p$ of being strain B.

   (a) What is the expected value of $\bar{X}$, where $\bar{X}$ is the sample mean of yields for 10 plants sampled from the population?

   (b) What is the variance of $\bar{X}$, where $\bar{X}$ is the sample mean of yields for 10 plants sampled from the population?

10. Suppose we are interested in estimating EX, but we can only observe data of the form $X_i + E_i$, where the X_i are drawn from the X population, and the E_i are measurement errors with expected value 0. Consider three situations (i) X_i and E_i are independent, (ii) X_i and E_i are positively correlated, (iii) X_i and E_i are negatively correlated. Given that our goal is to have the most precise estimate of EX, rank these three situations from the most favorable to the least favorable.

11. Suppose we make a single measurement x, and the quantity we are interested in is $g(x) = \sqrt{x}$. However the observed value of x is affected by measurement error of the form $x = x_0 + e$, where x_0 is the exact value that we intended to measure, and e is measurement error with mean 0 and standard deviation σ.

    (a) The delta method approximation for expected values is $Eg(X) \approx g(EX) + g''(EX)\text{var}(X)/2$. Derive an expression for the delta method approximation to the bias for this problem. Your answer will be a function of x_0 and σ.

(b) Is the approximate bias from part (a) positive or negative? Is it increasing or decreasing as a function of σ? Is it increasing or decreasing in magnitude as a function of σ? Is it increasing or decreasing as a function of x_0? Is it increasing or decreasing in magnitude as a function of x_0?

(c) The delta method approximation for the variance is $\text{var}g(x) \approx g'(EX)^2 \cdot \text{var}(X)$. Derive an expression for the delta method approximation to the variance for this problem. Your answer will be a function of x_0 and σ.

(d) Is the approximate variance from part (c) increasing or decreasing as a function of σ? Is it increasing or decreasing as a function of x_0?

12. (a) Suppose we observe a value X that is normally distributed with mean 1 and variance 2, and we observe a value Y that is normally distributed with mean 0 and variance 3. If X and Y are independent, what is the probability that we observe $X > Y$? Hint: derive an expression for $P(X - Y > 0)$.

    (b) Suppose we observe a value X that is normally distributed with mean 1 and variance 2, and we observe a value Y that is normally distributed with mean 0 and variance 3. If $\text{cor}(X, Y) = 1/2$, what is the probability that we observe $X > Y$?

    (c) Suppose X follows a normal distribution with mean 1 and variance 4. We have access to a standard normal quantile function `qnorm(p)`. Describe how we can use this to calculate the 90^{th} percentile of X.

13. Suppose we are interested in a quantity X, and our aim is to estimate EX as precisely as possible. Currently, we are only able to measure X with measurement error, so we observe $Y = X + E$, where E has expected value zero and is independent of X. The variance of E is $v = \sigma^2/10$, where σ^2 is the variance of X. We have done a pilot study with sample size n, and for our next study we are given two options:

    A Cut the variance of E in half, leaving the sample size at n.

    B Double the sample size, leaving the variance of E as it is.

    Which option is preferable, or are they equivalent?

14. Suppose we are carrying out a two-sample Z-test comparing a population X with mean μ_x and standard deviation σ_x to a population Y with mean μ_y and standard deviation σ_y. The size of the sample from the X population is m and the size of the sample from the Y population is n. State whether the power is "increasing," "decreasing," or "neither" as a function of each of the following values. "Neither" means that the power is increasing over part of its range and decreasing over another part of its range.

    (a) $\mu_x - \mu_y$

    (b) σ_x

(c) σ_y

(d) $p_x = m/(n+m)$

(e) $p_y = n/(n+m)$

(f) $m+n$

15. Suppose two researchers are comparing the expected value in a treated population (X) to a control population (Y) using the two sample Z-statistic. Both researchers are using 10 independent samples from each population.

    - Researcher A is studying a population in which $\mu_x - \mu_y = 1$, and $\sigma_x = \sigma_y = 1$.
    - Researcher B is studying a population in which $\mu_x - \mu_y = 2$, and $\sigma_x = \sigma_y = 3$.

    Which researcher has better power to reject a null hypothesis where $\mu_x = \mu_y$? Briefly explain your reasoning, but you do not need to derive a numerical value for either power.