

Statistics 403 Exam 1

October 14, 2008

Instructions:

- You may not use calculators, notes, books, formula cards, etc.
- 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. $P(Z < \dots)$, $P(Z > \dots)$), or in terms of R commands (e.g. `pnorm(...)`).
- Every problem is worth 15 points.
- Partial credit will be given. Express your answers clearly and show work where appropriate.

1. Suppose X_1, \dots, X_m are an independent sample from a population X with mean μ_X and standard deviation σ_X , and suppose Y_1, \dots, Y_n are an independent sample from a population Y with mean μ_Y and standard deviation σ_Y .

(a) Under the null hypothesis $\mu_X = \mu_Y$, what are the mean and standard deviation of $\bar{X} - \bar{Y}$?

Solution:

$$\begin{aligned} E(\bar{X} - \bar{Y}) &= E\bar{X} - E\bar{Y} \\ &= \mu_X - \mu_Y \\ &= 0. \end{aligned}$$

$$\begin{aligned} \text{var}(\bar{X} - \bar{Y}) &= \text{var}(\bar{X}) + \text{var}(\bar{Y}) \\ &= \sigma_X^2/m + \sigma_Y^2/n \end{aligned}$$

Therefore $\text{SD}(\bar{X} - \bar{Y}) = \sqrt{\sigma_X^2/m + \sigma_Y^2/n}$.

(b) Under an alternative hypothesis in which $\mu_X - \mu_Y = 1$, what are the mean and standard deviation of $\bar{X} - \bar{Y}$?

Solution: In this case, $E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y = 1$, and still $\text{SD}(\bar{X} - \bar{Y}) = \sqrt{\sigma_X^2/m + \sigma_Y^2/n}$.

2. Suppose we obtain the disease grade (a measure of severity of symptoms) for patients who are treated according to two different protocols, denoted A and B. Disease grade is scored quantitatively as 1, 2, or 3. The joint distribution of disease grades and treatment protocols in a particular hospital is as follows:

		Grade		
		1	2	3
A	0.2	0.1	0.3	
B	0.2	0.2	0.0	

- (a) Which protocol is used more commonly in this hospital?

Solution: $P(A) = 0.6$ and $P(B) = 0.4$, so protocol A is used more commonly.

- (b) What is the average grade for people who are on protocol B?

Solution: The conditional distribution of grades for people on protocol B is

		Grade		
		1	2	3
		0.5	0.5	0

Therefore the expected grade is $0.5 \cdot 1 + 0.5 \cdot 2 = 1.5$.

- (c) What is the probability that a patient is being treated with protocol B, given that his or her disease grade is 2?

Solution: The conditional probabilities of treatment protocol given grade are:

		Grade		
		1	2	3
A	0.5	0.33	1.0	
B	0.5	0.66	0.0	

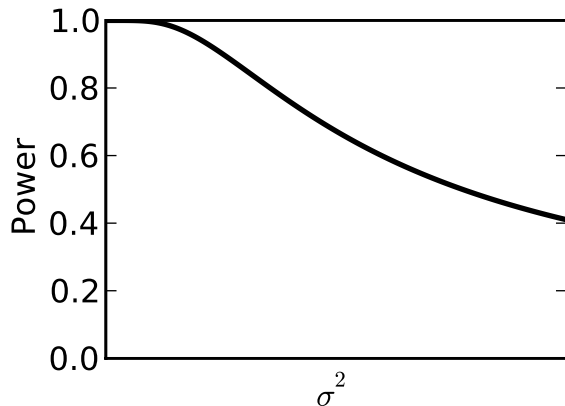
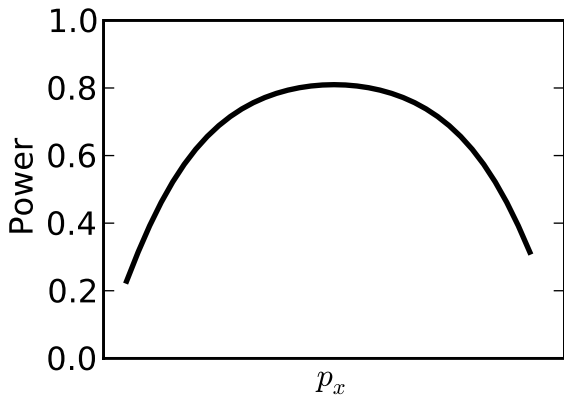
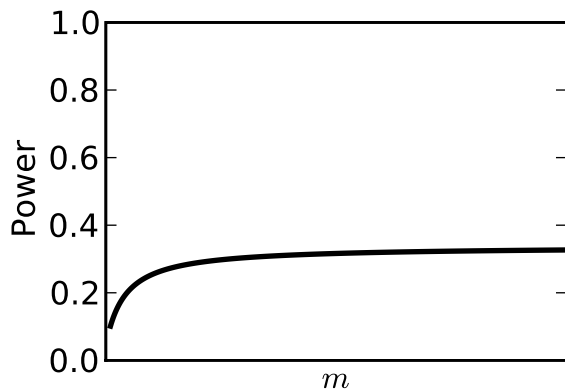
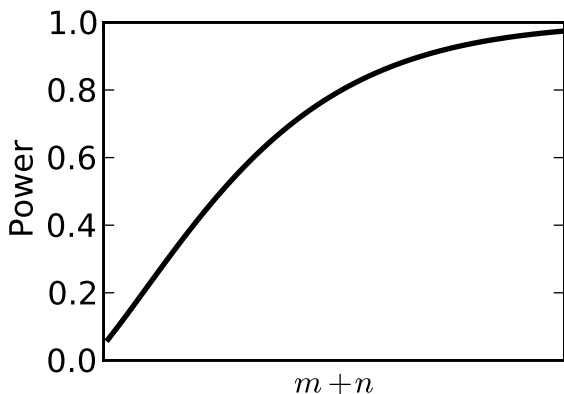
Therefore the probability that the treatment protocol is B given that the grade is 2 is $2/3$.

3. Suppose we plan to use a two-sample Z-test to assess whether the expected values of two populations, X and Y , are the same. As usual, we will sample independent values X_1, \dots, X_m from the X population and Y_1, \dots, Y_n from the Y population. The expected values of the X and Y populations are μ_x and μ_y , respectively, and the variances of the two populations are both σ^2 . For planning purposes we will assume that $\mu_x - \mu_y = 1$, and a two-sided test will be used.

To get a better sense for the power, plots were generated showing the power as a function of various things that affect the power. In each plot, one of the following quantities is varied along the horizontal axis, and the others are held fixed:

- σ^2 is varied while m and n are held fixed.
- $m + n$ is varied while σ^2 and $p_x = m/(m + n)$ are held fixed.
- m is varied while σ^2 and n are held fixed.
- $p_x = m/(m + n)$ is varied while $m + n$ and σ^2 are held fixed.

Label the horizontal axis in each of the four plots giving the quantity that is being varied. Each of the four quantities listed above should be used exactly once.



4. Suppose we sample X_1, \dots, X_m independently from a population X , and Y_1, \dots, Y_n independently from a population Y . We aim to assess whether $EX = EY$ using the two-sample Z-statistic.

(a) If $\bar{X} = 2$, $\bar{Y} = 1.2$, $\hat{\sigma}_x^2 = 1$, $\hat{\sigma}_y^2 = 1$, and $m = n = 10$, what is the p-value when the null hypothesis is $\mu_x = \mu_y$?

Solution: The test statistic is

$$\frac{2 - 1.2}{\sqrt{1/10 + 1/10}} = 4/\sqrt{5}$$

So the p-value is $P(Z < -4/\sqrt{5}) + P(Z > 4/\sqrt{5})$.

(b) Suppose $\mu_x - \mu_y = 1$, $m = n = 10$, and $\sigma_x = \sigma_y = 1$. We are given two options:

A Reduce the standard deviations (σ_x and σ_y) to 0.8

B Increase the sample sizes to get $m = n = 15$.

Determine which option is preferable in terms of power, or state that they are equivalent. Provide a brief description of your reasoning.

Solution: The power is greater for whichever of the two alternatives has a greater value of

$$T = \frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/m + \sigma_y^2/n}},$$

or of

$$\Delta = \frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/p_x + \sigma_y^2/(1 - p_x)}}.$$

Using the first expression, for option A we have

$$\frac{1}{\sqrt{0.8^2/10 + 0.8^2/10}} = \sqrt{1000/128}.$$

For option B we have

$$\frac{1}{\sqrt{1/15 + 1/15}} = \sqrt{15/2}.$$

With some arithmetic we can see that option A has a larger value, hence has more power.

5. Suppose we observe data X_1, \dots, 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” $\sigma = 3$ that is based on previously collected data of a similar type. However the truth is that $\sigma = 2$.

- (a) What is the actual coverage probability of our confidence interval?

Solution: The coverage probability is

$$\begin{aligned} P(-1.96 \leq \sqrt{n} \frac{\bar{X} - \mu}{3} \leq 1.96) &= P(-1.96 \cdot 3/2 \leq \sqrt{n} \frac{\bar{X} - \mu}{2} \leq 1.96 \cdot 3/2) \\ &= P(Z < 1.96 \cdot 3/2) - P(Z < -1.96 \cdot 3/2) \end{aligned}$$

- (b) What is the approximate width of our confidence interval?

Solution: Using the nominal standard deviation $\sigma = 3$, the width is approximately $2 \cdot 1.96 \cdot \sigma / \sqrt{n} \approx 12 / \sqrt{n}$.

- (c) What would be the approximate width of the confidence interval we would have gotten had we used the correct value for σ ?

Solution: Using the actual standard deviation $\sigma = 2$, the width is approximately $2 \cdot 1.96 \cdot \sigma \approx 8 / \sqrt{n}$.

6. Suppose we are studying anxiety in a particular population, and we obtain a measure X of anxiety on m family trios, each consisting of a mother, father, and child in the same family. The three individuals within a family have as their covariance matrix

$$\begin{pmatrix} 1 & 0.4 & 0.2 \\ 0.4 & 1 & 0.3 \\ 0.2 & 0.3 & 1.5 \end{pmatrix},$$

where the rows and columns are ordered as mother, father, and child. Assume that individuals in different families are independent.

- (a) Suppose we use the sample mean of all $3m$ people in our sample to estimate the mean level of anxiety in the population. What is the variance of this estimate?

Solution: The total of values in the within-family covariance matrix (given above) is 5.3. So if there are m families, the overall total of values in the covariance matrix is $5.3m$, so the variance of \bar{X} is $5.3m/(9m^2) = 5.3/(9m)$.

- (b) Describe how we could construct a 95% confidence interval for the mean level of anxiety in the population.

Solution: The 95% CI would be $\bar{X} \pm 1.96 \cdot \sqrt{5.3/(9m)}$.

7. Suppose in a particular industry, an input of X dollars of raw materials yields $1.1X^{1.5}$ dollars worth of product at the market value. We then select one company in this industry, and attempt to determine its raw material consumption X . Suppose the true raw material consumption is x_0 , and we observe $X = x_0 + E$, where E is measurement error with mean b and variance σ^2 . Our goal is to estimate the dollar value produced by this company.

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 , b , and σ^2 .

Hint: be careful when you go from the expected value to the bias, there is an important difference here compared to the other problems like this we have done.

Solution: Let $g(x) = 1.1x^{1.5}$, so $g'(x) = 1.1 \cdot 1.5 \cdot x^{0.5}$ and $g''(x) = 1.1 \cdot 1.5 \cdot 0.5/\sqrt{x}$. Thus the expected value of goods produced is around

$$1.1(x_0 + b)^{1.5} + 3.3\sigma^2/(8\sqrt{x_0 + b}).$$

So the bias is

$$1.1(x_0 + b)^{1.5} + 3.3\sigma^2/(8\sqrt{x_0 + b}) - 1.1x_0^{1.5}.$$