Statistics – Solutions

1. **D** – Of the choices, D is the only correct interpretation of a confidence interval.
2. **B** – The answer is the left endpoint of the confidence interval. Seen from the duality between confidence intervals and hypotheses testing.
3. **D** – Keep in mind that this is a one sided test. Z = 2.79, p = 0.003, reject at all reasonable levels of significance. Strong evidence that $\mu > 8$.
4. **E** – The answer is 1. Note that *f(x)* never takes on negative values, so no knowledge of calculus is necessary. Formally: P(X>0) = 1 – P(X≤0) = 1 - $\int_{-\infty}^{0} f(x)dx$ = 1 - $\int_{-\infty}^{0} 0dx$ = 1.
5. **C** – This can be seen after a simple integration; however, since calculus is not a prerequisite of statistics, we know that integration must not be required. We know that the area under a probability density function is always 1.
6. **D** – Because an easy test will have a high concentration of students scoring well, this distribution will be skewed to the left. In such a case, the mean will be below the median which will be below the mode. The only case in which this is true is choice D.
7. **A** – Definition of a Type II error.
8. **A** – A Chi-squared test for Goodness of Fit will be used, since we have an expectation and the observed counts are provided. The tricky part is determining the expected counts: There are possible combinations, listed as Player1, Player2. TT, TP, PT, TZ, ZT, PP, PZ, ZP, ZZ. Each of these should be equally likely to occur, with expected count 200/9 = 33.33. However, notice that the matchups involving two different races are actually repeated (for example, TP and PT are both Terran/Protoss matchups). So the expected counts of matchups are as follows: TT – 33.33, TP – 66.67, TZ – 66.67, PP – 33.33, PZ – 66.67, ZZ – 33.33. Now, using the standard Chi-squared formula of $(O-E)^2/E$, the Chi-squared value becomes 3.29, which, using 5 degrees of freedom, provides a p-value greater than 0.3. The null hypothesis, that each matchup is represented in the correct proportions, cannot be rejected.
9. **D** – P(X < (1377-1426)/σ) = .25. -49/σ = -.67, so σ=73.134 and σ^2 = 5,348.6.
10. **B** – P(X < (1500-1426)/73.13) – P(X < (1400-1426)/73.13) = .8441 - .3611 = 0.4830.
11. **C** – We first must create a new distribution for the herd of 200 cows: ~N(nμ, $\sqrt{n}$σ) = N(285,200, 1034.2). Then, we know that P(X<(x-285200)/1034.2) ≤ .05, so Z ≤ -1.645, meaning that x ≤ 283498.7. The largest whole number in that range is 283,498.
12. **C** – Either both D and E must work; or A, B, and C must all work. Since all components function independently, P(D and E) = p^2, P(A and B and C) = p^3. P(either scenario) = P(D,E) + P(A,B,C) = $p^2 + p^3$. When p = 0.6, this expression evaluates to 0.58.
13. **D** – There are three total covariates. Just because two of them are highly correlated does not imply that you can ignore one of them. Therefore, you must use all three.
14. **D** – The transformation is to enlarge the response variables Y, so we could take e^Y.
15. **B** – Definition of Simpson's Paradox.
16. **A** –This is a question of conditional probability. Construct the following table for help with organization.

	P	F	G	Total
D	.01*.3 = .003	.01*.4 = .004	.01*.3 = .003	.01
D^C	.99*.1 = .099	.99*.4 = .396	.99*.5 = .495	.99
Total	.102	.400	.498	1

Then, the two answers become clear:

A - P(D/P) = P(P&D)/P(P) = .003/.102 = 0.0294

17. **D** – Using the same table as in question 16, P(G) = 0.498
18. **E** – By De Morgan's identity, the complement of the intersection is the union of the complements, so the set is: $\{X < 3\}$ U $\{X > 5\}$. There are an infinite number of natural numbers that lie in this set.
19. **C** – The expression for the total commute time is $X_1 + X_2 + X_3 + X_4 + X_5$. Each coefficient is 1, and the variance of each day's time is 16. So the variance of the total weekly commute is $1^2*16 + 1^2*16 + 1^2*16 + 1^2*16 + 1^2*16 = 80$. Standard deviation = sqrt(80) = 8.94
20. **D** – One square in each chart is shaded, representing ¼ * ¼ = 1/16, or P(A)*P(B)=P(A and B).
21. **B** – Correlation is not a resistant measure (it is highly influenced by outliers). Transformation of either variable changes the correlation. The higher the magnitude of a correlation, regardless of its sign, the stronger the relationship. However, it was not described as a linear relationship. Therefore, part C is false. A correlation of 0 means there is no *linear* relationship but there could easily be a strong non-linear relationship.
22. **D** – The hypotheses are H_0: $\mu \leq 1/2$, H_A: $\mu > 1/2$. Since the sample size is small, perform a t-test. The observed Z-value is (.56-.5)/(0.05/(30^.5)) = 6.573. $P(t_{29} > 6.573) < 0.001$. Clearly, we should reject the null hypothesis at all reasonable significance levels.
23. **C** – The hypotheses are now H_0: $\mu_1 \leq \mu_2$, H_A: $\mu_1 > \mu_2$. The t-score is $t = \frac{X_1 - X_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}} =$ 1.80 for the given data. Using the conservative 29 degrees of freedom, the p-value lies between 0.01 and 0.05.
24. **B** – You could arrive 8:10-8:15 or 8:25-8:30 and have to wait under 5 minutes. This covers 10 of the 30 minutes, so the probability = 1/3
25. **B** – The time you arrive is a uniformly distributed random variable X where x is between 0 and 30. If $0<x<15$, Y_1 is a uniform random variable taking values 0-15. If $15<x<30$, Y_2 is again a uniform variable taking values 0-15. So Y (taking realized values a) is a uniform distribution spanning from 0 to 15. The expected value of such a distribution is E(Y) = (15-0)/2 = 7.5
26. **D** – Written in order the numbers are: 1 1 2 3 4 5 5 6 9. The 5-number summary (Min, 1st quartile, Median, 3rd quartile, Max) is: 1, 1.5, 4, 5.5, 9. 9-1 = 8, and 5.5-1.5 = 4 so the ratio is 2.
27. **A** – The observed counts are given in the chart, and we already have expected counts based on Natalie's hypothesis.
28. **A** – The null hypothesis for this test is H_0: $P_{Cantor} = P_{Gauss} = P_{Poincaré} = ¼$, $P_{Galois} = P_{Gödel} = P_{Perelman} = 1/12$. The observed/expected counts are:

	Cantor	Galois	Gauss	Gödel	Perelman	Poincaré
Observed	52	10	66	13	12	47
Expected	50	50/3	50	50/3	50/3	50

There are 6-1, or 5 degrees of freedom in this test. The Chi-square statistic yields 10.16, corresponding to a p-value of 0.07, so we cannot reject the null at a 5% significance level.
29. **B** – Changing hypothesis after seeing the raw data is defined as "data snooping" or "data fishing". It introduces bias and consequentially inflates the Type I error rate.
30. **B** – Let x=N(0,2) and y=N(0,2). $D^2/4 = (x^2+y^2)/4 = (x/2)^2 + (y/2)^2$. The distribution x/2 is represented by N(0,1) and y/2 is represented by N(0,1) as well. Thus we have $Z = X_1{}^2 + X_2{}^2$, where X_1 and X_2 are standard normal random variables, so $D^2/4$ is a Chi-square distribution with 2 degrees of freedom, which, by definition, has expected value 2.