Stevens Institute of Technology

SYS-601 Practice Exam #2 Solutions

2.1 Statistical Sampling

(a) Which of the following do random sampling methods seek to eliminate?

(A) Sampling error
(B) Correlation between samples
(C) Human bias
(D) X All of the above

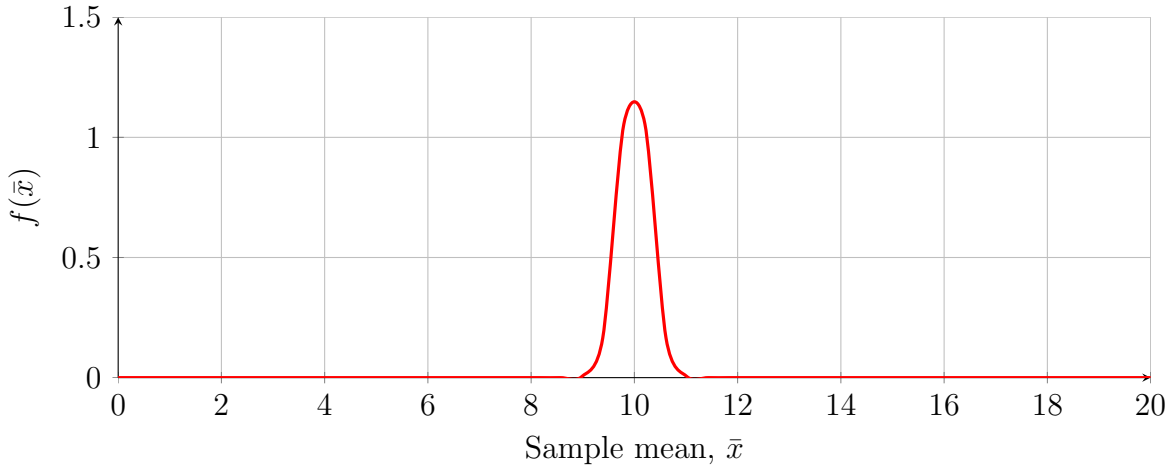
(b) Which situation generally recommends the use of Student's t distribution to compute confidence intervals on mean?

(A) Large number of samples
(B) X Unknown population std. deviation
(C) Normal population distribution
(D) None of the above

(c) If a random variable X has a normal distribution $f(x)$ with population mean $\mu = 10$ and standard deviation $\sigma = 3$, which of the following describes the distribution of the sample mean $\bar{X}$ for $N = 100$ samples?

(A) $\bar{X} \sim \text{normal}(\mu_{\bar{x}} = 10, \sigma_{\bar{x}} = 3)$
(B) X $\bar{X} \sim \text{normal}(\mu_{\bar{x}} = 10, \sigma_{\bar{x}} = 0.3)$
(C) $\bar{X} \sim \text{normal}(\mu_{\bar{x}} = 0, \sigma_{\bar{x}} = 1)$
(D) None of the above

(d) Sketch the distribution shape $f(\bar{x})$ for the sample mean in the space below.



(e) If a random variable X has a normal distribution $f(x)$ with population mean $\mu = 10$ and variance $\sigma^2 = 9$, which of the following describes the distribution of the sample variance s^2 for $N = 100$ samples?

(A) $11 \cdot s^2 \sim \text{chi2}(k = 9)$

(B) $s^2 \sim \text{chi2}(k = 9)$

(C) X $11 \cdot s^2 \sim \text{chi2}(k = 99)$

(D) None of the above

(f) Assume a normally-distributed population has standard deviation $\sigma = 3$. How many samples would be required to estimate the population mean with a 99% chance of being within 0.3 units of the true value?

$$N = \left(z_{\alpha/2} \cdot \frac{\sigma}{0.3}\right)^2 = \left(2.576 \cdot \frac{3}{0.3}\right)^2 = 25.76^2 = 664$$

2.2 Hypothesis Testing

(a) Which of the following statements is **incorrect**?

(A) The acceptable probability of making a Type I error is set by parameter α.

(B) A Type I error means a true null hypothesis was rejected.

(C) A Type II error means a false null hypothesis was not rejected.

(D) X Statistical methods minimize the chance of Type I and Type II errors.

(b) 3 PTS Which of the following is the best interpretation of a p-value?

(A) X Probability of a Type I error.

(B) Probability of a Type II error.

(C) Prob. of a true null hypothesis.

(D) Prob. of a false null hypothesis.

(c) Pokémon Go is a popular mobile game produced by Niantic. A large user community attempts to reverse-engineer various gameplay mechanics by collecting data and performing statistical analysis.

A dataset collected by Redditor CarolinaCapsFan in February 2017[1] investigates item drop rates. Out of $N = 1756$ samples, they report the sample drop rate of Potions to be $\bar{x} = 149/1756 = 0.0849$ and the sample standard deviation, approximated by a Bernoulli Trial, is $s = \sqrt{\bar{x}(1-\bar{x})} = \sqrt{0.0849 \cdot 0.9151} = 0.279$.

Perform a statistical test to check if the Potion drop rate during February 2017 was equal to a hypothesized value of $\mu^\star = 0.10$. Report the test statistic and p-value.

$$\text{Test statistic : } z = \frac{\bar{x} - \mu^\star}{\sigma/\sqrt{N}} = -2.278 \qquad p = 2 \cdot (1 - F_{\text{norm}}(|z|)) = 0.023$$

(d) Explain the meaning of results in (c) above to someone unfamiliar with statistics.

Results suggest with very low chance of being incorrect the true Potion drop rate is not 10% (it appears to be lower).

(e) A second dataset collected by Redditor Crawnic in August 2016[2] gathered $N = 9976$ samples, reporting a sample drop rate of Potions to be $\bar{x} = 760/9976 = 0.0762$ and the sample standard deviation is approximately $s = \sqrt{0.0762 \cdot 0.9238} = 0.265$.

Perform a statistical test to check if the Potion drop rate during August 2016 and February 2017 were equal. Report the test statistic and p-value. (*Hint: feel free to assume the sample standard deviations are close to true population standard deviations.*)

$$\text{Test statistic : } z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\sigma_1^2/N_1 + \sigma_2^2/N_2}} = 1.211 \qquad p = 2 \cdot (1 - F_{\text{norm}}(|z|)) = 0.226$$

[1] https://www.reddit.com/r/TheSilphRoad/comments/5vqjyc/pokestop_drop_rates_in_gen_2_lvl_31_7_days_511/

[2] https://www.reddit.com/r/TheSilphRoad/comments/4whv63/3207_pokestops_data/

(f) Explain the meaning of results in (e) above to someone unfamiliar with statistics.

Results do not provide sufficient evidence the two players' Potion drop rates are different from each other.

(g) Match each scenario with the best corresponding class of hypothesis test.

Check if the variance of one normally-distributed population is less than $\sigma^\star = 9$.	(i)	C	(A) z Test
Check if the means of two related populations are unequal.	(ii)	B	(B) Paired t Test
Check if the variances of two normally-distributed populations are unequal.	(iii)	D	(C) χ^2 Test
Check if the mean of one population is greater than $\mu^\star = 10$ using 100 samples.	(iv)	A	(D) F Test

2.3 Experimental Analysis

(a) Which of the following is **not** a key principle of rigorous experimental design?

(A) Randomization
(B) Replication
(C) X One factor at a time
(D) Blocking variables

(b) Why may a full factorial experiment **not** be desirable?

(A) Cannot detect interaction effects
(B) X Requires too many trials
(C) Results dependent on intuition
(D) Cannot isolate multiple factors

(c) On which statistical test is analysis of variation (ANOVA) based?

(A) X F test for difference in variances
(B) t test for difference in means
(C) χ^2 test for population variance
(D) z test for population mean

(d) Blogger Juan De Dios Santos collected data in Pokémon Go for nine days during August 2016 including information about item drops[3]. Out of $N = 886$ samples, he reports the sample drop rate of Potions to be $\bar{x} = 65/886 = 0.0733$. Does the addition of a third dataset help understand if the Potion drop rate varies across players? The three datasets are presented in the supplied file `potion_drops.csv`.

Calculate the following statistics using an appropriate ANOVA method. For ease of analysis, compiled datasets for the three Pokémaniacs are available via Canvas. Sorted samples only record 0 or 1 to determine if a Potion was dropped at a Pokéstop.

$$SSC = 0.126 \qquad MSC = 0.063$$

$$SSE = 898.7 \qquad MSE = 0.071$$

(e) Perform a statistical hypothesis test to check if at least one of the three players has a different Potion drop rate. Report the test statistic and p-value.

$$\text{Test statistic : } F_{2,12615} = \frac{MSC}{MSE} = 0.886 \qquad p = 1 - F_{\text{fdist}}(F_{2,12615}) = 0.412$$

(f) Explain the meaning of results to someone unfamiliar with statistics.

Results do not provide sufficient evidence the three players' Potion drop rates are different from each other.

[3] `https://github.com/juandes/PokemonGo9DaysAnalysis`

2.4 Regression Analysis

(a) Which statement regarding linear regression models is **incorrect**?

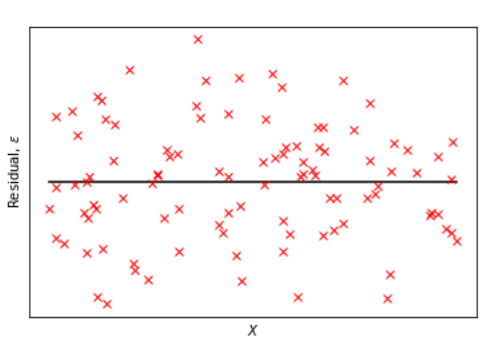
(A) Regression implies causality between independent and dependent variables.

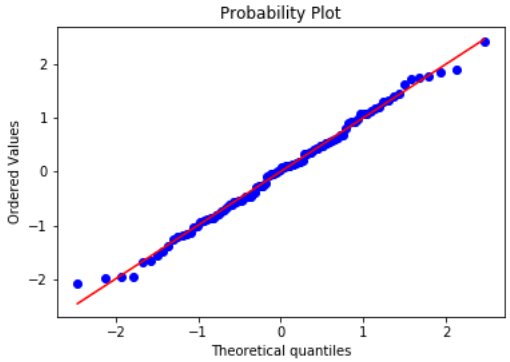
(B) X Forward selection and backward elimination methods will both produce the same multiple regression model.

(C) Unhealthy residuals indicate the linear assumption is invalid.

(D) Coefficient values are calculated to minimize the sum of squares of error between observations and predictions.

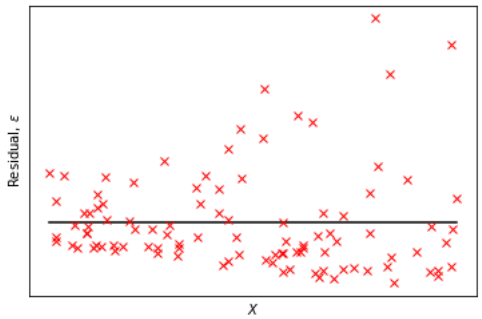
(b) Assess the health of the following simple regression model via its residual scatter (left) and probability (right) plots and mark **all** applicable properties.

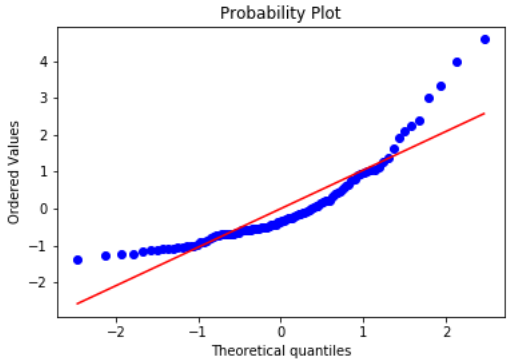




(A) X Homoscedasic (B) X Independent (C) X Normally distributed

(c) Assess the health of the following simple regression model via its residual scatter (left) and probability (right) plots and mark **all** applicable properties.





(A) Homoscedasic (B) X Independent (C) Normally distributed

(d) Blogger Juan De Dios Santos collected data about the Pokémon he caught in Pokémon Go for nine days during August 2016[4]. During this time period, he caught 60 Pidgey (a small bird) and recorded their combat power (CP), health points (HP), weight (kilograms), and height (meters).

The dataset is presented in the supplied file `Pidgey.csv`.

Develop a multiple regression model to predict Pidgey CP as a function of HP, weight, and height as follows:

$$\hat{y}_{\text{CP}} = \beta_0 + \beta_1 \cdot x_{\text{HP}} + \beta_2 \cdot x_{\text{weight}} + \beta_3 \cdot x_{\text{height}}$$

Report the fitted coefficient values, their statistical significance, and explain the meaning of results for each coefficient.

Coef.	Value	Statistically Significant?	Meaning of Results
β_0	-117.1	X Yes No	A Pidgey with zero HP, weight, and height would be expected to start at -117 CP.
β_1	6.6	X Yes No	Each additional HP increases CP by 6.6 points.
β_2	-1.8	Yes X No	Each kilogram of weight decreases CP by 1.8 points; however, this factor appears to be insignificant.
β_3	153.5	Yes X No	Each meter of height increase CP by 153.5 points; however this factor appears to be insignificant.

[4] `https://github.com/juandes/PokemonGo9DaysAnalysis`

2.5 Categorical and Non-parametric Statistics

(a) Which statement regarding non-parametric statistics is **incorrect**?

(A) They can accommodate ordinal, interval, or ratio sample data.
(B) They are based on sample rank rather than magnitude.
(C) X They are out-of-date compared to parametric statistics.
(D) They are not widely available in spreadsheet tools.

(b) Which statement regarding the chi-squared goodness of fit test is **incorrect**?

(A) It can accommodate any type of sample data.
(B) X Choosing how samples are binned into categories does not change results.
(C) No category should have zero expected observations.
(D) It can be used to check if data follow a mathematical distribution.

(c) Match each non-parametric statistic with the corresponding parametric version.

Friedman Test	(i)	E	(A)	t Test
Kruskal-Wallis Test	(ii)	D	(B)	Paired t Test
Mann-Whitney U Test	(iii)	A	(C)	Pearson's Correlation
Spearman's Rank Correlation	(iv)	C	(D)	One-way ANOVA
Wilcoxon Matched Pairs Test	(v)	B	(E)	Two-way ANOVA

(d) Some players of Pokémon Go are concerned that the current game mechanics reward players living in urban areas more than those in rural areas. The following (fictional) survey of 300 players shows the number of responses with negative, neutral, and positive sentiment for players living in urban and rural locations.

O_{ij}	Negative	Neutral	Positive
Urban	25	128	92
Rural	14	28	13

Record the expected number of observations below assuming player location and review sentiment are independent.

E_{ij}	Negative	Neutral	Positive
Urban	31.9	127.4	85.8
Rural	7.2	28.6	19.3

(e) Perform a statistical test to evaluate whether player location and review sentiment are independent. Report the test statistic and p-value.

$$\text{Test statistic} : \chi^2_2 = 10.536 \qquad p = 1 - F_{\text{chi2}}(\chi^2_2) = 0.005$$

(f) Explain the meaning of results to someone unfamiliar with statistics.

Results suggest with very low chance of being incorrect that player location and review sentiment are not independent factors.