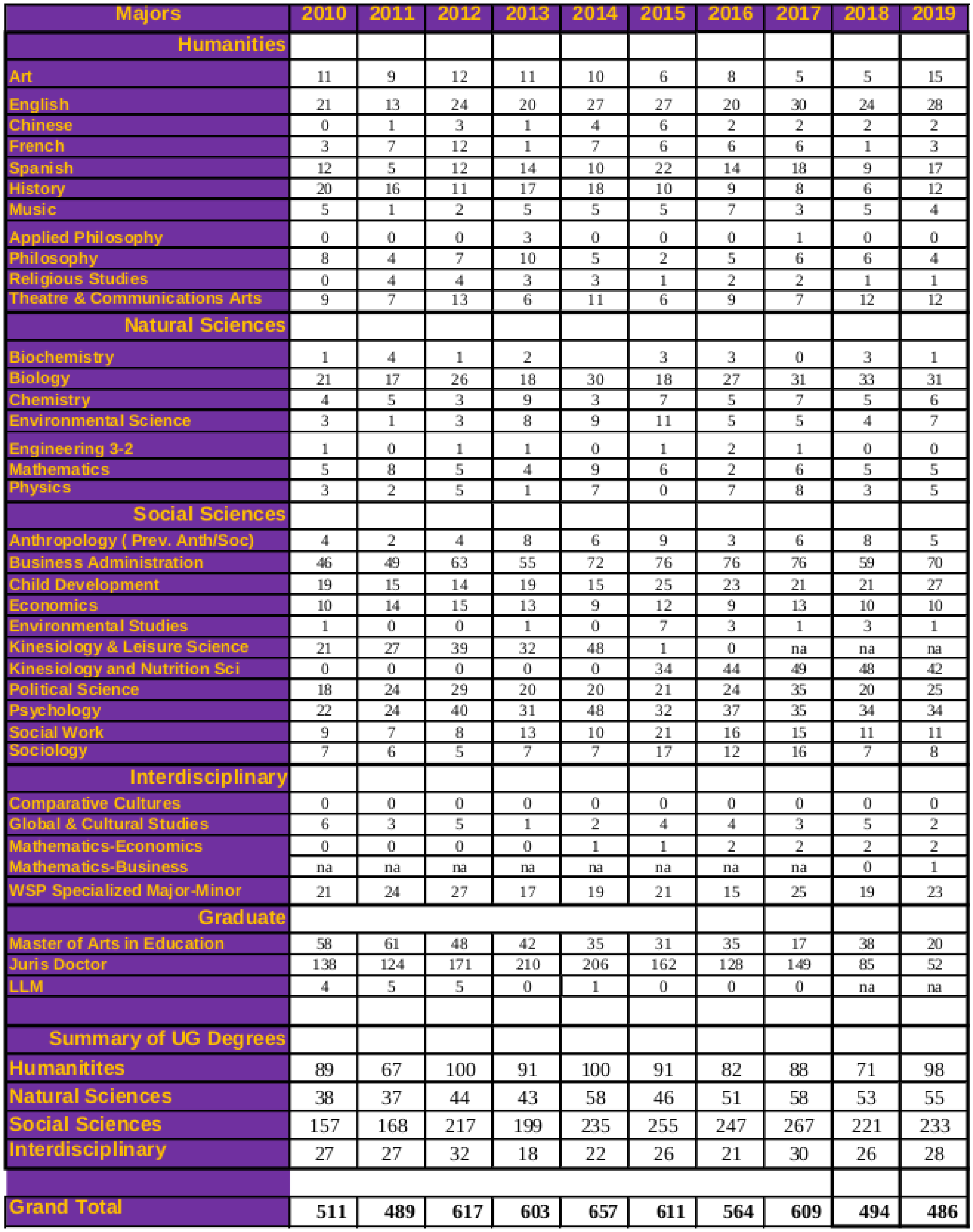
Study Guide for Final: Elementary Statistics

Prof. Jordan C. Hanson

August 10, 2020

1 Formula Area

1. Average/mean, definition 1: $\bar{x} = N^{-1}\sum_i x_i$
2. Median: the value below which are half of the frequencies. Half of the frequencies are also above this value.
3. Mode: the value corresponding to the highest frequendy.
4. The quartiles $Q1$, $Q2$, and $Q3$ are the values that separate the frequencies into four bins of equal frequency. $Q2$ is equal to the median. The IQR is $Q3 - Q1$.
5. The k-th percentile: the value below which k percent of the data is located. Formula: $i = (k/100)(n+1)$, where k is the percentile, n is the total number of data, and i is the integer location of the k-th percentile.
6. Finding the percentile of a data value: $(x + 0.5 * y)/n(100)$, where x is the number of data values below the given data value, y is the number of data values equal to the given one, and n is the total number of data values.
7. Average/mean, definition 2: $\bar{x} = \sum_i^M f_{r,i}x_i$, where x_i are the bin centers of a histogram, or the discrete random variable data values, and $f_{r,i}$ are the relative frequencies. For a discrete random variable, $f_{r,i}$ is replaced with $p(x)$, the probability distribution function.
8. Probabilities of mutually exclusive and independent events: if two events have probabilities p_1 and p_2, then the probability that event 1 AND event 2 occur is p_1p_2. The probability that event 1 OR event 2 occurs is $p_1 + p_2$.
9. The standard deviation s of a sample is
$$s^2 = \frac{1}{N-1}\sum_{i=1}^{M}(x_i - \bar{x})^2 \tag{1}$$
10. The mean and standard deviation of the binomial distribution are $\mu = np$ and $\sigma = \sqrt{npq}$, respectively.
11. Let the PDF of the *uniform distribution* be $p(x) = 1/(b+a)$. The mean and standard deviation of this PDF are $\mu = (b+a)/2$ and $\sigma = \sqrt{(b-a)/12}$, respectively.
12. Let the PDF of the *normal distribution* be $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-0.5(x-\mu)^2/\sigma^2)$. The mean and standard deviation of this PDF are μ and σ, respectively. We write $N(a,b)$ to refer to a normal distribution with mean a and standard deviation b.
13. The z-score of a result drawn from $N(\mu,\sigma)$ is $z = (x_i - \mu)/\sigma$. $P(|z| \leq 1) \approx 0.68)$, $P(|z| \leq 2) \approx 0.95)$, $P(|z| \leq 1) \approx 0.997)$.
14. **The central limit theorem** states that the means of samples of a population are distributed according to $N(\bar{x}, \sigma_x/\sqrt{n})$, if the sample size is n.
15. **The confidence interval** [a,b] may be constructed such that a fraction CL of all confidence intervals with the same properties will contain the population mean, μ. The number CL is called the *confidence level.*
16. Given the null hypothesis H_0 and an alternative hypothesis H_a, a **Type I error** is when we reject H_0 in favor of H_a, when H_0 is true.
17. Given the null hypothesis H_0 and an alternative hypothesis H_a, a **Type II error** is when we accept H_0 and reject H_a, when H_0 is false.



Majors	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Humanities										
Art	11	9	12	11	10	6	8	5	5	15
English	21	13	24	20	27	27	20	30	24	28
Chinese	0	1	3	1	4	6	2	2	2	2
French	3	7	12	1	7	6	6	6	1	3
Spanish	12	5	12	14	10	22	14	18	9	17
History	20	16	11	17	18	10	9	8	6	12
Music	5	1	2	5	5	5	7	3	5	4
Applied Philosophy	0	0	0	3	0	0	0	1	0	0
Philosophy	8	4	7	10	5	2	5	6	6	4
Religious Studies	0	4	4	3	3	1	2	2	1	1
Theatre & Communications Arts	9	7	13	6	11	6	9	7	12	12
Natural Sciences										
Biochemistry	1	4	1	2		3	3	0	3	1
Biology	21	17	26	18	30	18	27	31	33	31
Chemistry	4	5	3	9	3	7	5	7	5	6
Environmental Science	3	1	3	8	9	11	5	5	4	7
Engineering 3-2	1	0	1	1	0	1	2	1	0	0
Mathematics	5	8	5	4	9	6	2	6	5	5
Physics	3	2	5	1	7	0	7	8	3	5
Social Sciences										
Anthropology (Prev. Anth/Soc)	4	2	4	8	6	9	3	6	8	5
Business Administration	46	49	63	55	72	76	76	76	59	70
Child Development	19	15	14	19	15	25	23	21	21	27
Economics	10	14	15	13	9	12	9	13	10	10
Environmental Studies	1	0	0	1	0	7	3	1	3	1
Kinesiology & Leisure Science	21	27	39	32	48	1	0	na	na	na
Kinesiology and Nutrition Sci	0	0	0	0	0	34	44	49	48	42
Political Science	18	24	29	20	20	21	24	35	20	25
Psychology	22	24	40	31	48	32	37	35	34	34
Social Work	9	7	8	13	10	21	16	15	11	11
Sociology	7	6	5	7	7	17	12	16	7	8
Interdisciplinary										
Comparative Cultures	0	0	0	0	0	0	0	0	0	0
Global & Cultural Studies	6	3	5	1	2	4	4	3	5	2
Mathematics-Economics	0	0	0	0	1	1	2	2	2	2
Mathematics-Business	na	na	na	na	na	na	na	na	0	1
WSP Specialized Major-Minor	21	24	27	17	19	21	15	25	19	23
Graduate										
Master of Arts in Education	58	61	48	42	35	31	35	17	38	20
Juris Doctor	138	124	171	210	206	162	128	149	85	52
LLM	4	5	5	0	1	0	0	0	na	na
Summary of UG Degrees										
Humanitites	89	67	100	91	100	91	82	88	71	98
Natural Sciences	38	37	44	43	58	46	51	58	53	55
Social Sciences	157	168	217	199	235	255	247	267	221	233
Interdisciplinary	27	27	32	18	22	26	21	30	26	28
Grand Total	**511**	**489**	**617**	**603**	**657**	**611**	**564**	**609**	**494**	**486**

*Number of majors awarded exceeds the number of degrees awarded due to double majors

Figure 1: Information regarding awarded Whittier College degrees.

1. Consider Fig. 1. (a) Create a *time-series* of the ratio of social science degrees to total degrees awarded versus time. (b) Create the same ratio, but for natural sciences. (c) Create a *pareto chart* for the fraction of degrees in humanities, natural sciences, social sciences, and humanities awarded in 2019. (d) What mean number of physics degrees awarded per year, averaged over the past 10 years? What is the standard deviation?

2. Consider Fig. 1. (a) Create a histogram of the total degrees awarded each year, and determine the (b) quartiles of the data.

3. Suppose the number of biology degrees awarded at Whittier College awarded per year is described by $N(20, 5)$. (a) Suppose one year, we award 30 degrees in biology. What is the z-score of this result? (b) What is the probability of this result occuring by random chance? (c) What are the mean and standard deviation of the actual biology degrees across the last decade, according to Fig. 1? (d) What is the z-score of the most recent number of biology degrees?

4. Consider the following statistical claim: "If we select a random Whittier College undergraduate student, the probability we will guess his or her major is 1/37, because there are 37 undergraduate majors." What is wrong with this claim, statistically?

5. Suppose we repeatedly measure the mean number of psychology degrees found in a randomly selected sample of size $n = 100$ students. (a) The mean is 10 and the standard deviation is 3. What is the standard error in the mean, accordint to the CLT? (b) Construct a confidence interval that contains the true mean number of psychology degrees (per 100 students) that contains the true mean 95% of the time.

6. An office manager is interested in the mean number of emails each worker sends during his or her work day. A survey of 30 employees is taken. The mean from the sample is 6.0 with a sample standard deviation of 4.0. (a) Construct the 95 percent confidence interval for the population mean, given the data. (b) Graph the distribution of mean number of emails per day. (c) Mark the confidence interval on the graph.

7. When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is the drug is unsafe. What is the Type II Error?

   - A: To conclude the drug is safe when in, fact, it is unsafe.
   - B: Not to conclude the drug is safe when, in fact, it is safe.
   - C: To conclude the drug is safe when, in fact, it is safe.
   - D: Not to conclude the drug is unsafe when, in fact, it is unsafe.

8. Suppose a new medicine moves ahead with human trials. When people are given a placebo dose, 10 percent of them are "cured." The fraction of patients cured with the new medicine is 20 ± 5 %. Suppose we construct a null hypothesis H_0: "If the fraction of patients cured is greater than or equal to the result corresponding to two standard deviations above the placebo result, then the drug is effective." (a) Should we reject or confirm the null hypothesis? At what significance level is this result? (b) Suppose there was a problem with the data, and the true rate of cure is actually 15 ± 5 %. What has happened to the significance level of the drug trial?