# STAT 515 fa 2023 Exam II

Karl Gregory

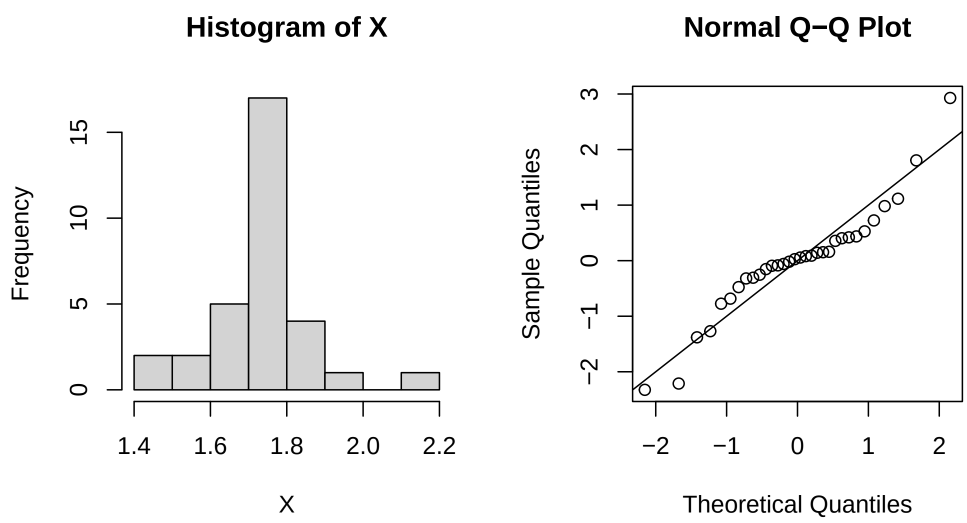
- Do not open this exam until told to do so.
- You may have two handwritten sheet of notes out during the exam.
- You have 75 minutes to work on this exam.
- You may NOT use any kind of calculator.
- If you are unsure of what a question is asking for, do not hesitate to ask me for clarification.
- *Good luck, and may the odds be ever in your favor!*

$$\hat{p}_n \pm z_{\alpha/2} \cdot \sqrt{\hat{p}_n(1-\hat{p}_n)/n} \qquad \bar{X}_n \pm t_{n-1,\alpha/2} \cdot S_n/\sqrt{n}$$

$$Z_{\text{test}} = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \qquad T_{\text{test}} = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$$

A $z$-table and a $t$-table are appended to this exam.

1. The length and diameter of each acorn in a sample of 32 live oak acorns was measured (in this class!!). Suppose the ratio of the length to the diameter is of interest. The 32 length-to-diameter ratios in the sample had mean $\bar{X}_n = 1.74$ and standard deviation $S_n = 0.138$. The figure below shows a histogram and a normal QQ plot of the length-to-diameter ratios.




(a) Explain the purpose of the normal QQ plot and give your interpretation of this one.

The purpose to check whether the data appear to have been drawn from a population with a Normal distribution. There is some snake-like behavior of the points in this QQ plot, so it does NOT appear that the data were drawn from a Normal distribution.

(b) A 95% confidence interval for the mean length-to-diameter ratio of live oak acorns is constructed with an formula like this:

$$(i) \pm (ii)\frac{(iii)}{(iv)}.$$

Give the numbers to plug in for $(i)$, $(ii)$, $(iii)$, and $(iv)$.

The interval is given by

$$1.74 \pm 2.0395\frac{0.138}{\sqrt{32}},$$

where $2.0395 = t_{32-1,0.025}$

(c) Of the three intervals $(1.69, 1.79)$, $(1.70, 1.78)$, and $(1.67, 1.81)$, one is the 90% confidence interval, one is the 95% confidence interval, and one is the 99% confidence interval based on these data for the mean length-to-diameter ratio of live oak acorns. Which interval is the 90% confidence interval?

The 90% confidence interval will be the narrowest one, so it is the interval $(1.70, 1.78)$.

(d) Suppose one wished to test whether the mean length-to-diameter ratio of live oak acorns was the golden ratio 1.618. Give the hypotheses of interest, using $\mu$ to denote the mean length-to-diameter ratio of the live oak acorn population.

We are interested in testing $H_0$: $\mu = 1.618$ versus $H_1$: $\mu \neq 1.618$.

(e) The test statistic for testing the hypothesis is computed with a formula like this:

$$T_{\text{test}} = \frac{(i) - (ii)}{(iii)/\sqrt{(iv)}}.$$

Give the numbers to plug in for $(i)$, $(ii)$, $(iii)$, and $(iv)$.

We would compute

$$T_{\text{test}} = \frac{1.74 - 1.618}{0.138/\sqrt{32}}.$$

(f) The test statistic value is $T_{\text{test}} = 5.011806$. Give your conclusion about the golden ratio hypothesis using significance level $\alpha = 0.01$.

In order to reject $H_0$ at significance level $\alpha = 0.01$, the test statistic must exceed, in absolute value, the threshold $t_{32-1,0.005} = 2.7440$. Since $T_{\text{test}} = 5.011806 > 1.7440$, we reject the null hypothesis at significance level $\alpha = 0.01$. We therefore conclude that the golden ratio does not apply to the length-to-diameter ratio of live oak acorns.

2. In a survey of 38 students (in this class!!), 11 reported that they had a houseplant. Let's regard the 38 students as a random sample of USC students.

(a) The Wald-type 95% confidence interval for the proportion of USC students with a houseplant is constructed with a formula like this:

$$(i) \pm (ii)\sqrt{\frac{(iii)}{(iv)}}.$$

Give the numbers to plug in for $(i)$, $(ii)$, $(iii)$, and $(iv)$.

The Wald-type interval is computed as

$$11/38 \pm 1.96\sqrt{\frac{11/38(1 - 11/38)}{38}}.$$

(b) For the Agresti-Coull interval (which has much better performance), we add two "successes" and two "failures" to the data set and recompute the Wald-type interval. Give the numbers $(i)$, $(ii)$, $(iii)$, and $(iv)$ such that

$$(i) \pm (ii)\sqrt{\frac{(iii)}{(iv)}}$$

gives the Agresti-Coull interval for the proportion of USC students with a houseplant.

> The Agresti-Coull interval is computed as
>
> $$13/42 \pm 1.96\sqrt{\frac{13/42(1-13/42)}{42}},$$
>
> where $1.96 = z_{0.025}$.

(c) The 95% Agresti-Coull interval is $(0.170, 0.450)$. Give an interpretation of this interval.

> We are 95% confident that the proportion of USC students who have a houseplant is between 0.170 and 0.450.

(d) Suppose you wish to more accurately estimate the proportion of USC students with houseplants. Specifically, suppose you wish to estimate it within 1 percentage point with 99% confidence. Give an expression for the sample size required (you do not have to simplify your expression).

> We need sample size no smaller than $(2.5758)^2 \cdot 11/38(1-11/38)/(0.01)^2$.

(e) The required sample size from part (d), if we use the survey data to make a guess at the population proportion, comes out to $n = 13{,}647$, which you decide is too large. How can you change your specifications in part (d) to make the required sample size smaller?

> We can allow a larger margin of error—that is we can aim to estimate the true proportion to within 2 percentage points instead of 1 percentage point—or we can reduce the confidence level from 99% to, for example, 95%. Either change would result in a smaller required sample size.

(f) Suppose your botany professor claims that no more than 10% of USC students have houseplants. Give the null and alternate hypotheses for testing his claim.

> The hypotheses of interest are $H_0$: $p \leq 0.10$ versus $H_1$: $p > 0.10$.

(g) For testing the hypotheses in part (f), suppose you compute

$$Z_{\text{test}} = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} = 2.407$$

using the survey data. Give the corresponding $p$-value.

The $p$ value is the area under the standard Normal pdf to the right of the value 2.407. We can obtain this from the $z$-table as $0.5000 - 0.4920 = 0.0080$.

(h) Give your conclusion about the claim of the botany professor in part (f). Use significance level $\alpha = 0.05$.

Since the $p$-value is less than 0.05, we reject the null hypothesis. Therefore we conclude that the proportion of USC students with a houseplant is greater than 0.10.

3. Suppose the weights of bananas on sale at your grocery store have a Normal distribution with a mean of 135 grams and a standard deviation of 15 grams.

(a) Give the probability that a randomly selected banana weighs between 120 and 150 grams.

We have $P(120 < X < 150) = P((120 - 135)/15 < Z < (150 - 135)/15) = P(-1 < Z < 1) = 2(0.3643) = 0.6826$.

(b) Give the probability that the mean of the weights of 9 randomly selected bananas falls between 120 and 150.

We have $P(120 < \bar{X}_9 < 150) = P((120 - 135)/(15/3) < Z < (150 - 135)/(15/3)) = P(-3 < Z < 3) = 2(0.4987) = 0.9974$

(c) Give an explanation for why there is a difference between the answers to parts (a) and (b).

The reason the answers are different is that in part $(b)$ the probability is computed for a mean of a sample rather than for a single sampled value. The variance of a sample mean is $1/n$ times the variance of a single sampled value.

4. Students taking a survey (in this class!!) were asked to weigh their keychains and record the weight in grams. Thirty-five students weighed their keychains. The mean weight was 84.71 grams. Consider the three sets of hypotheses:

| (1) | (2) | (3) |
|---|---|---|
| $H_0$: $\mu \geq 70$ | $H_0$: $\mu = 70$ | $H_0$: $\mu \leq 70$ |
| $H_1$: $\mu < 70$ | $H_1$: $\mu \neq 70$ | $H_1$: $\mu > 70$ |

When the survey data are used to test these sets of hypotheses, the tests result in the $p$-values below; match each $p$-value to one of the hypotheses (1), (2), or (3).

(a) The $p$-value 0.0434.

The data, having mean 84.71, supports the alternate hypotheses of (2) and (3). Since the $p$-value (iii) is twice this one, this one must correspond to the one-sided set of hypotheses in (3), and the $p$-value in (iii) must correspond to the two-sided set of hypotheses in (2).

(b) The $p$-value 0.9566.

The data support the null hypothesis of (1), so for this set of hypotheses we expect a large $p$-value.

(c) The $p$-value 0.0868.

This corresponds to the two-sided set of hypotheses in (2)