STATS 242 HW 3

February 4, 2025

Number of late days: 0; Collaborators: Alex Wa, Derek Gao

1.

a) We detail our devised two possible test statistics:

(1) **Test Statistic T_1: Z-Statistic**
For this test statistic, we want to check whether the expected number of male children in each family is actually $12(0.5) = 6$, what we would expect under H_0 if $X_i \sim \text{Bin}(12, 0.5)$. In other words, $H_0 : \mu = 6$ and $H_1 : \mu \neq 6$. Because the variance of each X_i is known under H_0, we can use the Z-statistic as our test statistic:

$$T_1 = \frac{\sqrt{6115}}{\sigma}\bar{X}$$

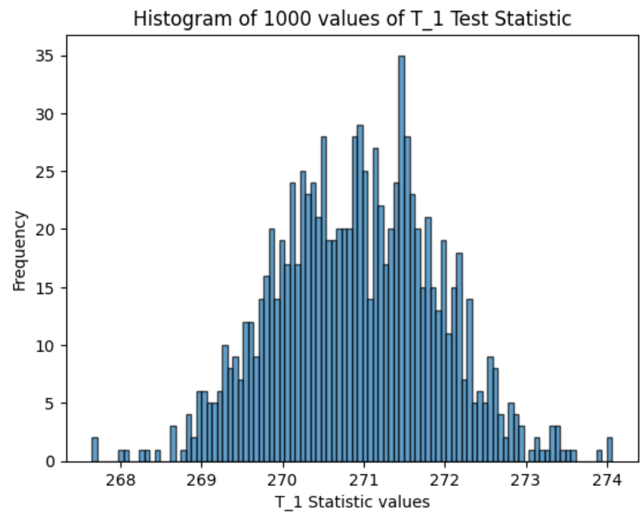
In my code, we compute the T_1 statistic for this data to be **260.48**.

(2) **Test Statistic T_2: Sample Variance**
For this test statistic, we test if sample variance is actually equal to $12(0.5)(0.5) = 3$, as it would be under H_0 where the observed male child frequences follow a $\text{Bin}(12, 0.5)$ distribution. Our test statistic is the sample variance:

$$T_2 = \frac{\Sigma_{i=1}^{6115}(X_i - \bar{X})^2}{6115 - 1}$$

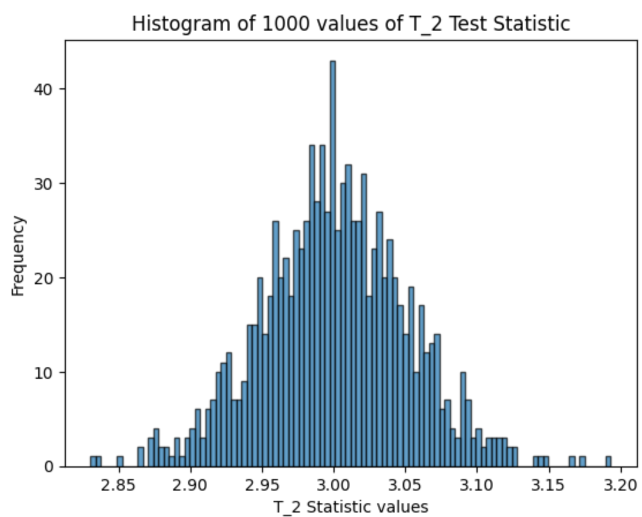
In my code, we compute the T_2 statistic for this data to be **3.490**.

b) (1) **Test Statistic T_1**
Below is our histogram for the null distribution of the T_1 test statistic.



For this test statistic at the α=0.05 significance level, the upper-α point is **268.60** and the lower-α point which is **268.03**, which is less than the observed T_1 statistic for this data (**260.48**). Thus, we reject H_0.

(2) **Test Statistic T_2**
Below is our histogram for the null distribution of the T_2 test statistic.



For this test statistic at the α=0.05 significance level, the upper-α point and lower-α point are given by **2.875** and **2.834** respectively. Because the observed T_2 statistic for this data (**3.490**) is greater than the upper-α point, we reject H_0.

```
    # %%
"""Run all imports"""
```

```python
import numpy as np
import matplotlib.pyplot as plt
import math

NUM_MALES = np.array([i for i in range(0, 12 + 1)])
NUM_FAMILIES = np.array([7, 45, 181, 478, 829, 1112, 1343, 1033, 670, 286, 104, 24, 3])
VARIANCE = 12 * 0.5 * (1 - 0.5)

assert np.sum(NUM_FAMILIES).item() == 6115

def t1_statistic(num_males, num_families):
    mean_num_males = np.sum(num_families * num_males) / np.sum(num_families)
    return np.sqrt(6115 / VARIANCE) * mean_num_males

def t2_statistic(num_males, num_families):
    mean_num_males = np.sum(num_families * num_males) / np.sum(num_families)
    return np.sum(num_families * (num_males - mean_num_males) ** 2) / (6115 - 1)

print(f"On this data, our T_1 statistic is {t1_statistic(NUM_MALES, NUM_FAMILIES)}")
print(f"On this data, our T_2 statistic is {t2_statistic(NUM_MALES, NUM_FAMILIES)}")
# %%
def get_data_for_one_simulation():
    # for 6115 families, sample the number of children they have based on Bin(12, 0.5)
    num_males = NUM_MALES
    num_families = np.zeros_like(NUM_FAMILIES)
    num_males_6115 = np.random.binomial(12, 0.5, 6115)
    for i, num_male in enumerate(num_males):
        num_families[i] = np.sum(num_males_6115 == num_male)
    assert np.sum(num_families) == 6115
    return num_males, num_families


N_SIMULATIONS = 1000

def get_statistics_simulation(statistic_func):
    test_statistics = []
    for _ in range(N_SIMULATIONS):
        num_males, num_families = get_data_for_one_simulation()
        test_statistics.append(statistic_func(num_males, num_families))
```

```python
    return np.array(test_statistics)

"""
Run 1000 simulations to appromimate null distribution of T_1 and T_2
 Statistics.
"""

t1_statistics = get_statistics_simulation(t1_statistic)
t2_statistics = get_statistics_simulation(t2_statistic)
"""
Plot histogram of these statistics.
"""
plt.hist(t1_statistics, bins=100, edgecolor='black', alpha=0.7)

# Add labels and title
plt.xlabel('T_1 Statistic values')
plt.ylabel('Frequency')
plt.title('Histogram of 1000 values of T_1 Test Statistic')
plt.show()

plt.hist(t2_statistics, bins=100, edgecolor='black', alpha=0.7)
plt.xlabel('T_2 Statistic values')
plt.ylabel('Frequency')
plt.title('Histogram of 1000 values of T_2 Test Statistic')
# %%
"""
For the T_1 Statistic, can we reject?
"""
SIGNIFICANCE_LEVEL = 0.05
OBSERVED_T1_STATISTIC = t1_statistic(NUM_MALES, NUM_FAMILIES)
t1_critical_value_upper_end = np.percentile(t1_statistics, 0.95)
t1_critical_value_lower_end = np.percentile(t1_statistics, 0.05)
if (t1_critical_value_upper_end <= OBSERVED_T1_STATISTIC): print("Can
 reject null hypothesis for T_1 test statistic.")
if (t1_critical_value_lower_end >= OBSERVED_T1_STATISTIC): print("Can
 reject null hypothesis for T_1 test statistic.")

"""
For the T_2 Statistic, can we reject?
"""
SIGNIFICANCE_LEVEL = 0.05
OBSERVED_T2_STATISTIC = t2_statistic(NUM_MALES, NUM_FAMILIES)
t2_critical_value_upper_end =  np.percentile(t2_statistics, 0.95)
t2_critical_value_lower_end =  np.percentile(t2_statistics, 0.05)
if (t2_critical_value_upper_end <= OBSERVED_T2_STATISTIC): print("Can
 reject null hypothesis for T_2 test statistic.")
```

```
if (t2_critical_value_lower_end >= OBSERVED_T2_STATISTIC): print("Can
↪  reject null hypothesis for T_2 test statistic.")

# %%
```

2.

a) Note that because H_0 is true, then we can assume the true distribution of to be $Z \sim \mathcal{N}(0,1)$. To find the distribution of $P = 1 - \Phi(z)$, we compute its CDF below:

$$P(P \leq p) = P(1 - \Phi(Z) \leq p)$$

Note that by the Universality of the Uniform, $\Phi(Z) \sim \text{Unif}(0,1)$. Therefore, $1 - \Phi(Z) = 1 - \text{Unif}(0,1) = \text{Unif}(-1,0) + 1 = \text{Unif}(0,1)$. Let us define $U \sim \text{Unif}(0,1)$. Putting this all together, we have:

$$P(P \leq p) = P(1 - \Phi(Z) \leq p) = P(U \leq p) = F_U(p)$$

where F_U is the CDF of U. Because r.v. P has the same CDF as $U \sim \text{Unif}(0,1)$, we can conclude $P \sim \text{Unif}(0,1)$ if H_0 is true.

b) If H_1 was true, then we would expect to observe greater values of test statistic $Z \implies \Phi(Z)$, which is strictly non-decreasing, would be greater $\implies P = 1 - \Phi(Z)$ would be smaller. Thus, if H_1 was true, we would expect to observe smaller p-values. As shown in part (a), $P \sim \text{Unif}(0,1)$ so $P_{H_0}(P \leq \alpha) = F_U(\alpha) = \alpha$. Therefore, if we were to use P as the test statistic, we would reject H_0 when $P \leq \alpha$.

3.

a) Let us define $Z_1, \ldots, Z_n$ to all be i.i.d $\mathcal{N}(0,1)$ distributions. We can define distribution $U_n = \frac{Z_1^2 + \cdots + Z_n^2}{n}$, or equivalently that $U_n \sim \frac{1}{n} \cdot \chi_n^2$. We first compute $\mathbb{E}[Z_i]^2 = \text{Var}[Z_i] + \mathbb{E}[Z_i] = 1 + 0 = 1$. By the Weak Law of Large Numbers, since $\mathbb{E}[Z_i] = 1$ and Z_i has finite variance, $U_n \to \mathbb{E}[Z_i]$ or $U_n \to 1$ in probability as $n \to \infty$.

Let us now define function $g(z) = \sqrt{z}$, where g is continuous for $(0, \infty)$. By the Continuous Mapping Theorem, distribution $g(U_n) = \sqrt{U_n} \to \sqrt{1}$, or $\sqrt{U_n} \to 1$ in probability as $n \to \infty$.

b) The t-distribution with n degrees of freedom is given by:

$$T_n = \frac{\frac{\sqrt{n}}{\sigma}\bar{X}}{\sqrt{\frac{1}{n}\chi_n^2}}$$

where $\bar{X} = \frac{X_1+\cdots+X_n}{n}$ and $\sigma = \sqrt{\mathrm{Var}(X_i)}$. As shown in part (a), $\sqrt{\frac{1}{n}\chi_n^2} \to 1$ in probability as $n \to \infty$. Furthermore, by CLT as $n \to \infty$, $\bar{X} \sim \mathcal{N}(0, \sigma^2)$ and thus $\frac{\sqrt{n}}{\sigma}\bar{X} \to \mathcal{N}(0,1)$. Because $\sqrt{\frac{1}{n}\chi_n^2} \to 1 \neq 0$, we can apply Slutsky's Lemma. By Slutsky's Lemma, $T_n = \frac{\frac{\sqrt{n}}{\sigma}\bar{X}}{\sqrt{\frac{1}{n}\chi_n^2}} \to \frac{\mathcal{N}(0,1)}{1}$ or $T_n \to \mathcal{N}(0,1)$ as $n \to \infty$.

4.

(a) $T \sim t_1$ is given by $\frac{X}{\sqrt{\frac{1}{1}\chi_1^2}} = \frac{X}{\sqrt{Y^2}}$, where X, Y are i.i.d $\mathcal{N}(0,1)$. By the Continuous Mapping Theorem, given $g(x) = \sqrt{x^2} = |x|$ then $g(Y)$ will converge to $|Y|$ and so T is given by $\frac{X}{|Y|}$.

We can define $\frac{X}{|Y|}$ as the following:

$$\frac{X}{|Y|} = \begin{cases} \frac{X}{Y} & \text{if } Y > 0 \\ -\frac{X}{Y} = \frac{X}{-Y} & \text{if } Y < 0 \end{cases}$$

Because Y and $-Y$ have the same distributions (because Y is symmetric about the origin), both the $\frac{X}{Y}$ and $\frac{X}{-Y}$ have the same distribution (which we can call U). Because then we have $\frac{X}{|Y|} = U \cdot P(Y > 0) + U \cdot P(Y < 0) = \frac{U+U}{2} = U$, we can conclude $\frac{X}{|Y|} \sim U$ or $\frac{X}{|Y|} \sim \frac{X}{Y}$[1]

b) Let us define $g(x,y) = (\frac{x}{y}, y)$. Because g is bijective, its inverse can be given by $g^{-1}(t,u) = (tu, u)$. We can use the change-of-variables formula to compute the joint PDF $f_{T,U}(t,u)$[2]:

$$f_{T,U}(t,u) = f_{X,Y}(g^{-1}(t,u)) \cdot |\det \begin{pmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \\ \frac{\partial y}{\partial t} & \frac{\partial y}{\partial u} \end{pmatrix}|$$

$$f_{T,U}(t,u) = f_{X,Y}(tu, u) \cdot |\det \begin{pmatrix} \frac{\partial (tu)}{\partial t} & \frac{\partial (tu)}{\partial u} \\ \frac{\partial u}{\partial t} & \frac{\partial u}{\partial u} \end{pmatrix}|$$

$$f_{T,U}(t,u) = f_{X,Y}(tu, u) \cdot |\det \begin{pmatrix} u & t \\ 0 & 1 \end{pmatrix}|$$

$$f_{T,U}(t,u) = |u| f_{X,Y}(tu,u) = |u| f_X(tu) f_Y(u) = |u| \frac{1}{2\pi} e^{-\frac{[(tu)^2+u^2]}{2}} = |u| \frac{1}{2\pi} e^{-\frac{u^2(t^2+1)}{2}}$$

We now compute f_T:

$$f_T(t) = \int_{-\infty}^{\infty} f_{T,U}(t,u)du = \int_{-\infty}^{\infty} \frac{|u|}{2\pi} e^{-\frac{u^2(t^2+1)}{2}} du = \int_{-\infty}^{0} \frac{-u}{2\pi} e^{-\frac{u^2(t^2+1)}{2}} du + \int_{0}^{\infty} \frac{u}{2\pi} e^{-\frac{u^2(t^2+1)}{2}} du$$

[1] Note that $P(Y < 0) = P(Y > 0) = 0.5$ because Y is a normal distribution and is therefore symmetric about the origin.

[2] Note that because X and Y are independent, $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.

Note that these two integrals are equivalent (we can use a change of variables $dv = -u$ in the first one) and so we can combine them:

$$f_T(t) = \frac{1}{\pi}[\int_0^\infty ue^{-\frac{u^2(t^2+1)}{2}} du]$$

Using a u-substitution for $v = \frac{u^2(t^2+1)}{2}$, we have:

$$f_T(t) = \frac{1}{\pi}\int_0^\infty \frac{1}{t^2+1}e^{-v}dv = -\frac{1}{\pi(t^2+1)}[e^{-v}]\Big|_0^\infty = -\frac{1}{\pi(t^2+1)}[0-1] = \frac{1}{\pi(t^2+1)}$$

Given the PDF of T we now can compute its expectation:

$$\mathbb{E}[T^2] = \int_{-\infty}^{\infty} \frac{t^2}{\pi(t^2+1)}dt = \frac{1}{\pi}[\int_{-\infty}^{\infty} 1 - \frac{1}{t^2+1}dt] = \frac{1}{\pi}[t - \arctan(t)]\Big|_{-\infty}^{\infty}$$

which diverges to ∞. Thus we have $\mathbb{E}[T^2] = \infty$.