CLASS 9 – MULTIPLE TESTING AND CHI-SQUARED TESTS

1. **Iran Election Revisited**
2. **Chi-squared Tests**
3. **Evaluating Fraud**
4. **Conclusions**
5. **Appendix I – Extra Information on Chi-squared Distribution**

Sources: DeGroot and Schervish (DGS) Chapters 10

1. IRAN ELECTION REVISITED

In Problem Set 4 I asked you to examine the last digit from reported vote results in Iran and conduct a hypothesis test. Variations of "digit" tests like this have been used to evaluate electoral and general accounting and financing fraud.[1]

In setting up your hypothesis, you should have selected a significance level (α) before your test which would determine the critical region / rejection region for your test.

QUESTION: If you set $\alpha = 0.05$, what should the Type I error rate from your test be?

+/- 1.96 σ

QUESTION: What was the approximate p-value from your hypothesis test, and what was your conclusion?

0.009

[1] See for example Berber and Scacco 2012 and here.

THE ACTUAL TYPE I ERROR RATE FROM OUR PROCEDURE:

46 % = probability that we'd reject the Null-Hypothesis even if it is true
(say it is not true, when it is in fact true)
=
claiming an event is rare when, in fact, it is not

QUESTION: Why is our Type I error rate so much larger than our significance level?

ELECTION DATA FROM THE 116 PROVINCES:

(1)	(2)	(3)
Last Digit	Frequency of Digit	Observed Fraction
1	11	9.48%
2	8	6.90%
3	9	7.76%
4	10	8.62%
5	5	4.31%
6	14	12.07%
7	20	17.24%
8	17	14.66%
9	13	11.21%
0	9	7.76%
Total	116	100%

Using the Central Limit Theorem might lead to error…
Pick 7 as last digit to talk about possible fraud, pick another digit (such as 1) to show no fraud has occurred.

Idea of testing for 7 arrived after looking at the data: changed the way of random testing process.

With 7, process of picking the maximum. Problem is it is only looking at the distribution of digit 7 ("is it 7 or not?"), ignoring all other information.

Maybe the number 7 "takes away" from other digits in a non-uniform way (ex. look at 5, 3, 0).

KEY THEMES

Suppose you were to pick just one single proportion to test the randomness of the final digits for the Iranian vote counts. Which would you pick if you wanted to show there was fraud in the election?

In the problem set, you tested whether one fraction p was equal to 0.10 (the value we assumed for p under our null hypothesis)

- But we have ten different fractions
- We don't want to test just one – we want to test them all!

We need to use **multiple testing** approaches to test several hypotheses at once, otherwise we risk inflating our Type I error rate.

QUESTION: Suppose you run two independent null hypothesis tests at the 5% significance level. Assume for both tests the null hypothesis is in fact true. What is the probability that at least one of the tests rejects the null?

Test A | 5%
|
|
Test B | 5%

Probability that test A fails to reject: 95%
Probability that test B fails to reject: 95%
Probability that at least one of them succeeds to reject:
1–(0.95)(0.95) = 9.75%

QUESTION: Suppose you run 100 independent null hypothesis tests at the 5% significance level. Assume for all tests the null hypothesis is in fact true. How many times do you expect to reject the null?

100 * 5% = 5

It might be okay, if we were running only one analysis.
When running multiple analyses, Type I errors get inflated.

There are many forms of multiple testing, some you have already seen:

- F-tests in the regression context (courtesy of PP455)

The F-test = Cumulative test of whether all coefficients are equal to 0

- In the special case where the error term has no heteroskedasticity and no autocorrelation the F-statistic takes the following form:

$$F(q, n-k) = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k)} \quad (9)$$

  - $SSR_u = \sum(\hat{u}_i)^2$ in the unrestricted model
  - $SSR_r = \sum(\hat{u}_i)^2$ in the restricted model
  - n is the number of observations
  - q is the number of additional coefficients estimated in the unrestricted model
  - k is the number of coefficients in the unrestricted model

QUESTION: You conduct an F test to see if the means of four groups are the same. You get an F statistic of 2.03. Given that the F statistic is greater than 1.96 you should reject the null hypothesis of equality of means.

1.96 is for a Normal Distribution

Changing to an F distribution, the test statistic cannot be compared to 1.96!
Cannot do that with something that is not Normal-distrib.

2. CHI-SQUARED TESTS

BASICS OF CHI-SQUARED TESTS

χ^2

Can be used for:
- goodness of fit
- comparison between distributions

- Generally used for variables with more than two discrete values
- Test of entire distributions rather than single proportions
- Most common uses of Chi-Squared tests:
    1. Comparison of a single distribution to a known population **[Goodness of fit]**
        - Is the distribution of births by day of the week uniform?
    2. Comparison of two distributions **[Independence, Association, and/or Homogeneity]**
        i. Are two different samples likely to have been drawn from the same distribution (without explicitly defining what that distribution is)

Today's example focuses on the first use of chi-square tests: Goodness of fit

STEPS IN A CHI-SQUARED TEST

There is nothing special about this (or in general any other) hypothesis test. The procedure is the same as for tests we have seen, but our test statistic and distribution will be different.

1. State the null hypothesis
2. Set a significance level (α)
3. Calculate the **test statistic** from the sample
4. Use the **distribution of the test statistic** to calculate a p-value
5. **Reject** or **fail to reject** the null hypothesis

Steps:
- Write a null hypothesis,
- Calculate α,
- Calculate test statistic,
- Use distribution to find p-value,
- (Fail to) Reject the Null

QUESTION: What is our null hypothesis for this test?

H_0: p_0 = 0.1
p_2 = 0.1
…
p_q = 0.1

H_1: "the null is false" = at least one of the values (one of the p_q's) is not equal to 1.

α = 0.05

TEST STATISTIC

For a goodness of fit test, if we have *k* different groups (e.g. groups could be the different digits), we can calculate the relevant test statistic by examining the sum of the squared deviations of the observed counts from the expected counts (scaled by the expected counts). Mathematically, we can calculate the "chi-squared" (χ^2) test statistic as:

Test carried out in terms of (raw) counts

$$\chi^2 = \sum_{k} \frac{(observed\ count_k - expected\ count_k)^2}{expected\ count_k}$$

This test statistic will follow a chi-squared distribution with *k*-1 degrees of freedom.

(1)	(2)	(3)	(4)	(5)
Last Digit	Frequency of Digit	Observed Fraction	Expected Fraction	Expected Count
1	11	9.48%	10%	11.6
2	8	6.90%	⋮	⋮
3	9	7.76%	⋮	⋮
4	10	8.62%	⋮	⋮
5	5	4.31%	⋮	⋮
6	14	12.07%	⋮	⋮
7	20	17.24%	⋮	⋮
8	17	14.66%	⋮	⋮
9	13	11.21%	⋮	⋮
0	9	7.76%	v	v
Total	116	100%	100%	

0.1 * 116
⋮
v
…

(1) Last Digit	(2) Observed Count	(5) Expected Count	X^2 Stat term
1	11	11.6	(11 – 11.6)^2 / 11.6 = 0.013
2	8	11.6	1.12
3	9	11.6	0.58
4	10	11.6	0.22
5	5	11.6	3.76
6	14	11.6	0.50
7	20	11.6	6.08
8	17	11.6	2.51
9	13	11.6	0.17
0	9	11.6	0.58
Total	116	116	X^2 Stat = 15.55

(driven by 3 big categories: 7, 5, 8)

F distribution is not a Normal distribution, it is something else

```
# chisquare function from scipy.stats package
observed_digits = np.array([11, 8, 9, 10, 5, 14, 20, 17, 13, 9])
chisquare(f_obs = observed_digits)
```

✓ 0.0s

```
Power_divergenceResult(statistic=15.551724137931034, pvalue=0.07685369980326165)
```

Statistical interpretation of p-value = 0.077:
probability of seeing a test statistic at least as big as seen (15.55), if the null hypothesis is true.

Need to determine whether the observed test statistic is large or small: it depends on the specific case chi-squared distribution. Those distributions change shaped based on how many categories there are.
Think about where the extreme values live in the distribution… only in the RIGHT TAIL (rejection region only in the right-hand side – the distribution is positive only, contrary to the N-distrib).

3. FROM P-VALUES TO FRAUD

QUESTION: In conditional probability notation, how do we write the reported p-value above?

When we calculate the p-value: p [Data | H_0], or more extreme

QUESTION: In conditional probability notation, how do we write the value we would like to know to assess fraud?

"What is the probability of fraud, given our data" = p [Fraud | Data],
by itself it does not demonstrate fraud.

We would want to do a full Baysian analysis.

p [Fraud | Data] can be found using the analysis showed above.

4. CONCLUSIONS

Beware of **cherry-picking**

Be careful with **multiple testing**

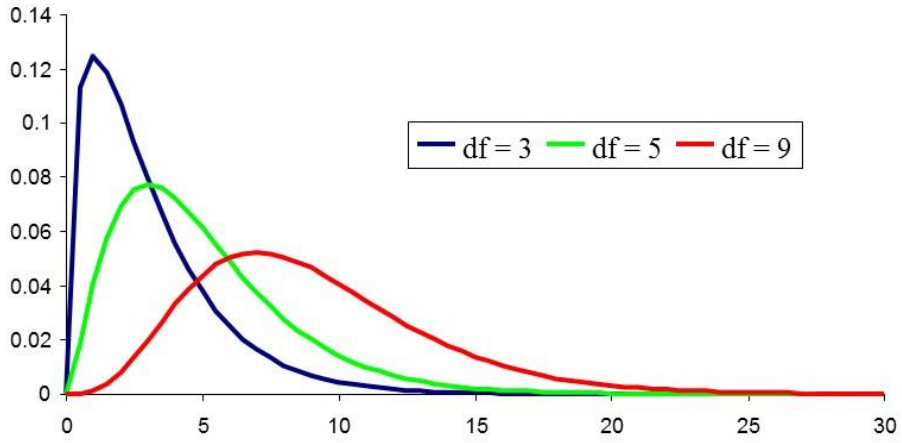
- Testing multiple relationships individually could cause you to **over-reject** the null hypothesis
    - The more tests you conduct, the more likely you are to reject the null hypothesis in one test just by chance
    - A single test which aims to examine all of this information combats this "multiple testing" problem
- Testing multiple relationships individually could also cause you to **under-reject** the null hypothesis
    - If every individual test is close to the rejection region, but fails individually, a collective test could succeed in rejecting the null hypothesis

p-values are interpreted relative to the hypothesis test we ran. If we want to extend that to a different conclusion, we need to be ready to justify the connection, using Bayes' Rule if necessary.

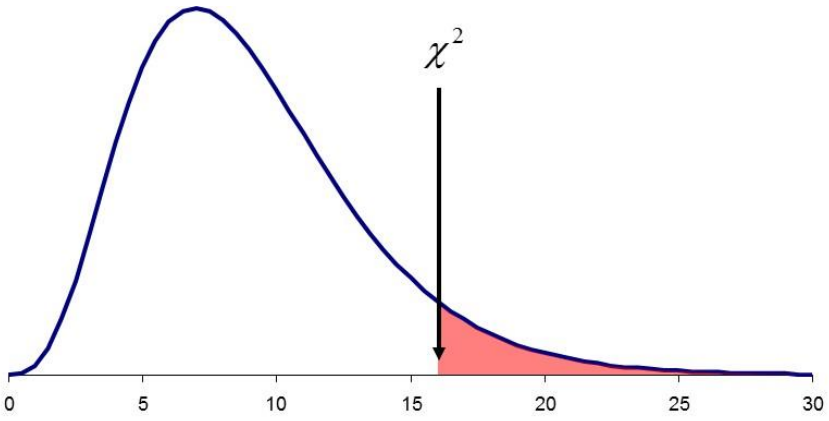
APPENDIX I – EXTRA INFORMATION ON CHI-SQUARED TESTS

DEGREES OF FREEDOM IN χ^2 DISTRIBUTION

- Chi-squared requires a "degrees of freedom" (df) = Number of categories – 1
- When there are more categories, there are more numbers added together to form the Chi-squared test statistic



- In the case of the Iranian votes, df = 9 [stemming from 10 different digits]
- Our test statistics was ~ 15.5, corresponding to the below critical region



TESTING THE HYPOTHESIS THAT $H_0: q = 0.10$ IN THE 2009 IRANIAN ELECTIONS

- Steps 1-2 – State the **null hypothesis** $(H_0) = H_0: q = 0.10$ and **significance level** (α) $\alpha = 0.05$
- Steps 3-5 – **Assuming H_0 is true**, determine likelihood of observing an estimate as extreme or more extreme than the one you drew (i.e. calculate p-value)

Using Sampling Distribution of Estimator

Estimate: $\hat{q} = 0.172$

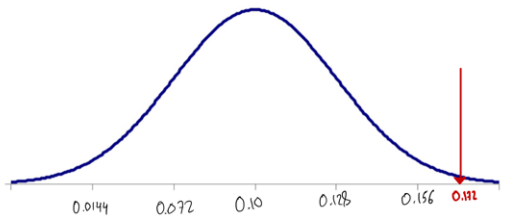
Sampling Distribution (assuming Ho is true):

Shape: Approx. Normal

Mean: 0.10

Standard Deviation: $\sqrt{\frac{q(1-q)}{n}} = \sqrt{\frac{0.10(1-0.10)}{116}} = 0.028$

Graphically:



Calculate p-value:

p-value=0.01

Since p-value<α, reject $H_0: q = 0.10$

Using Distribution of Test Statistic

Value of Test Statistic: $z = \frac{\hat{q}-q_0}{SE} = \frac{0.172-0.10}{0.028} = 2.57$

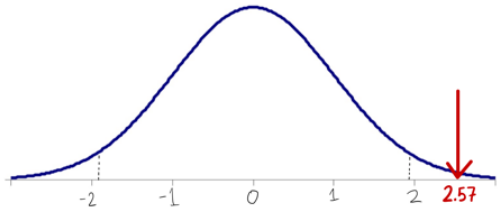
Distribution of Test Statistic (assuming Ho is true)

Shape: Approx. Normal

Mean: 0

Standard Deviation: 1

Graphically:



Calculate p-value:

p-value=0.01

Since p-value<α, reject $H_0: q = 0.10$

The following scenarios illustrate examples of where individual t-tests and a chi-squared test can lead to different conclusions.

Hypothetical scenario #1 – Under-reject

Final Digit	Actual Count	Pred. Count	p-value of indiv. test	X^2 Stat
1	16	11.5	0.162	1.76
2	16	11.5	0.162	1.76
3	16	11.5	0.162	1.76
4	16	11.5	0.162	1.76
5	16	11.5	0.162	1.76
6	7	11.5	0.162	1.76
7	7	11.5	0.162	1.76
8	7	11.5	0.162	1.76
9	7	11.5	0.162	1.76
0	7	11.5	0.162	1.76
Total	115		X^2 Stat:	17.61 **p-value = 0.04**

- None of the individual tests rejects the null hypothesis
- The Chi-squared test rejects the null hypothesis
- In words:
  - Individual tests suggest no issue with H_o
  - Chi-squared test rejects H_o

Hypothetical scenario #2 – Over-reject

Final Digit	Actual Count	Pred. Count	p-value of indiv. test	X^2 Stat
1	2	11.6	**0.003**	7.94
2	13	11.6	0.665	0.17
3	13	11.6	0.665	0.17
4	13	11.6	0.665	0.17
5	12	11.6	0.901	0.01
6	13	11.6	0.665	0.17
7	13	11.6	0.665	0.17
8	14	11.6	0.458	0.50
9	10	11.6	0.620	0.22
0	13	11.6	0.665	0.17
Total	116		X^2 Stat:	9.69 ***p-value* = 0.38**

- One of the individual tests rejects the null hypothesis
- The Chi-squared test fails to reject the null hypothesis
- In words:
  - A single individual tests rejects H_o
  - Chi-squared test fails to reject H_o

- Bottom line is that you should do a test that considers **all of the information at once** and not do individual tests in a piecemeal manner.