Quantitative Economics Lecture 6 - Statistical Inference II : Testing Multi-sample and Multiple Hypotheses

Richard Povey

University of Oxford

2^{nd} May 2018

richard.povey@hertford.ox.ac.uk, richard.povey@st-hildas.ox.ac.uk

With thanks to Ian Crawford, Margaret Stevens, Vanessa Berenguer-Rico.

Outline of Lecture

- Here we first introduce the concept of a p-value.
- Then we cover some more complex examples of hypothesis testing, where we have *multiple samples* and/or *multiple hypotheses* to simultaneously test.
- Sometimes it is possible to construct a single t-test statistic for the hypothesis, in which case the procedure is essentially the same as covered in Lecture 5.

Outline of Lecture

- Sometimes we will have *more than one* t-statistic (2 in all of the cases covered in this lecture, but in regression analysis later on in the course you will use multiple hypothesis tests with *more than 2* degrees of freedom).
- Here we can run either a chi-squared test or an F-test (we will see that they are closely related).
- The procedure for a chi-squared test is to square each of the k t-stats and add them together in order to construct a $\chi^2(k)$ variable.
- We also give a simple example of an F-test in this lecture, but you will see more useful examples later in the course.

Outline of Lecture

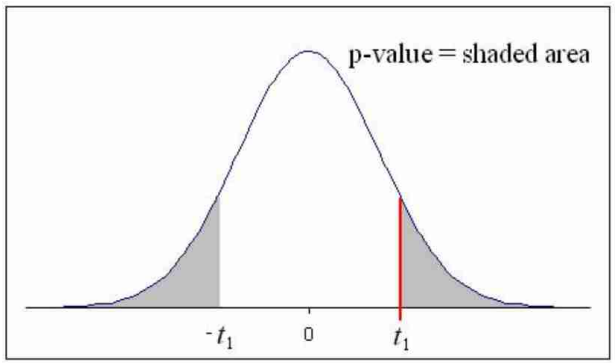
A few examples of multi-sample and/or multiple hypothesis tests:

- **Comparing the means** in two populations using two distinct samples (a t-test).
- **Testing more than one population parameter at once** - In this lecture we give the example of testing that the means in two distinct populations are *both* equal to a particular value (a $\chi^2(2)$ test).
- **Testing for a normal distribution** in a population using a single sample (simultaneously testing that $Skew(X) = 0$ and $Kurt(X) = 3$) (a $\chi^2(2)$ test).

The p-value

- A very useful concept when hypothesis testing is the p-value.
- This is the probability of the "tail" of the distribution left at a particular realised value of a test statistic.
- For a two-sided t-test of a single hypothesis with realised value t_1, the p-value is:

$$p = 2\Phi(-|t_1|) = 2\left(1 - \Phi(|t_1|)\right)$$



Computer output normally provides the two-sided p-value by default.

- For a one-sided t-test, with $H_a : \mu > \mu_0$: $p = 1 - \Phi(t_1)$
- For a one-sided t-test, with $H_a : \mu < \mu_0$: $p = \Phi(t_1)$

Hypothesis Testing for the Difference Between Means

Suppose we want to compare the mean of *average hourly earnings* for male and female workers in the CPS2004.

Separate the sample into male and female sub-samples:

	Female	Male
Mean	15.3586	17.7726
Standard Error	0.1339	0.1361
Median	13.9423	15.5289
Standard Deviation	7.7100	9.3036
Sample Variance	59.4436	86.5566
Minimum	2.0979	2.1368
Maximum	57.6923	61.0577
Count	3313	4673

$$\hat{\mu}_f = \bar{Y}_f = 15.3586$$
$$s_f = 7.7100$$
$$\hat{\mu}_m = \bar{Y}_m = 17.7726$$
$$s_m = 9.3036$$

Let d be the difference between the population means: $d = \mu_m - \mu_f$.

We could estimate d using: $\hat{d} = \bar{Y}_m - \bar{Y}_f$.

To test hypotheses about the true value of d we need to know the distribution of $\hat{d}$.

Hypothesis Testing for the Difference Between Means

Assume that $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means of a random variable Y for two *independent* random samples.

If the samples are large: $\bar{Y}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$ and $\bar{Y}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right)$

Applying results for the distribution of $aX + bY$:

$$\begin{aligned} \mathrm{E}(\hat{d}) &= \mathrm{E}(\bar{Y}_1) - \mathrm{E}(\bar{Y}_2) = \mu_1 - \mu_2 \\ \mathrm{Var}(\hat{d}) &= \mathrm{Var}(\bar{Y}_1) + \mathrm{Var}(\bar{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad \text{(independence)} \end{aligned}$$

and since both $\bar{Y}_1$ and $\bar{Y}_2$ are normal, so is $\hat{d}$: $\hat{d} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

Then we can *estimate* the variance of $\hat{d}$ by using the sample variances, and obtain the t-statistic:

$$\mathrm{se}(\hat{d}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad t = \frac{\hat{d} - d}{\mathrm{se}(\hat{d})}$$

Hypothesis Testing for the Difference Between Means

As before, the t-statistic $t = \dfrac{\hat{d} - d}{\text{se}(\hat{d})}$ has a standard normal distribution when the sample is large.

Then we can carry out tests such as:

$$\left.\begin{array}{l} H_0 : d = d_0 \\ H_a : d \neq d_0 \end{array}\right\} \quad \text{Under H}_0, \;\; t = \frac{\hat{d} - d_0}{\text{se}(\hat{d})} \sim N(0,1). \quad \text{Reject } H_0 \text{ at 5\% level if } |t| > 1.96$$

- Amongst US employees aged 25-34 in 2004, do men earn more than women?

We could use a one-sided test here:

$$\left.\begin{array}{l} H_0 : d = 0 \\ H_a : d > 0 \end{array}\right\} \quad \text{Under H}_0, \;\; t = \frac{\hat{d}}{\text{se}(\hat{d})} \sim N(0,1). \quad \text{Reject } H_0 \text{ at 5\% level if } t > 1.64.$$

$$d = 17.7726 - 15.3586 = 2.4141 \qquad \text{se}(\hat{d}) = \sqrt{\frac{9.3036^2}{4673} + \frac{7.7100^2}{3313}} = 0.19096$$

$$t = \frac{2.4141}{0.19096} = 12.64. \qquad \text{We reject } H_0 \text{ at the 5\% level. The p-value is 0.}$$

How to interpret this? Do male and female workers have the same education level?

Hypothesis Testing for the Difference Between Proportions

To find out whether the proportion of women in the labour force is the same in 1992 and 2004:

$$\text{In 2004}: \begin{cases} \hat{p}_1 = 0.4149 \\ n_1 = 7986 \end{cases} \quad \text{and in 1992}: \begin{cases} \hat{p}_2 = 0.4262 \\ n_2 = 7602 \end{cases}$$

and for $\hat{d} = \hat{p}_1 - \hat{p}_2$ we want to test: $H_0 : d = 0$ against $H_a : d \neq 0$.

Under H_0, both populations have the same mean, $p_1 = p_2 = p$, and the same variance $p(1-p)$:

$$\hat{d} \sim N\left(0, \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}\right) = N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

So for the standard error of $\hat{d}$, we can estimate p in the *pooled* sample:

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} \quad \Rightarrow \quad \text{se}(\hat{d}) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\hat{d} = -0.0113, \quad \hat{p} = 0.4204, \quad \hat{p}(1-\hat{p}) = 0.2437, \quad \text{se}(\hat{d}) = 0.00791$$

Then: $t = \frac{-0.0113}{0.00791} = -1.429$. We can't reject the null at the 5% level; the p-value is 0.1531.

Understanding the F-test

- Just as a t-test is asymptotically equivalent to a z-test, an F-test is asymptotically equivalent to a chi-squared test. This means that when you (or a computer for you) run a t-test or F-test, if n is high then this is equivalent to a z-test or chi-squared test.
- When you do regression analysis later in the course, you will see that the usual F-test statistic for a multiple hypothesis takes the form:

$$f = \frac{\left(\frac{RSS_r - RSS_u}{q}\right)}{\left(\frac{RSS_u}{n-k}\right)} = \left(\frac{RSS_r - RSS_u}{RSS_u}\right)\left(\frac{n-k}{q}\right)$$

- Here q is the number of restrictions being made and k is the number of independent variables (here *including* the *constant* or *intercept* variable).
- RSS stands for **residual sum of squares**: $RSS = \sum_{i=1}^{n}\left[\left(X_i - \hat{X}\right)^2\right]$, where $\hat{X}$ is the **fitted/predicted value** of X_i, which in the case of hypothesis testing about the population mean is simply $\hat{\mu}_X = \bar{X}$.

Relationship between F-test and t-test

How does this work for the sample mean ($q = 1$ and $k = 1$)?

$$\textit{Restricted model: } \hat{X} = 0 \Longrightarrow RSS_r = \sum_{i=1}^{n} \left[(X_i - 0)^2\right] = \underline{\sum_{i=1}^{n} \left[X_i^2\right]}$$

$$\textit{Unrestricted model: } \hat{X} = \bar{X} \Longrightarrow RSS_u = \sum_{i=1}^{n} \left[(X_i - \bar{X})^2\right]$$

$$\Longrightarrow RSS_u = (n-1)\bar{S}^2 = \sum_{i=1}^{n} \left[X_i^2 - 2X_i\bar{[}X] + \bar{X}^2\right]$$

$$= \sum_{i=1}^{n} \left[X_i^2\right] - 2\bar{X} \sum_{i=1}^{n} \left[X_i\right] + n\bar{X}^2$$

$$= \underline{\sum_{i=1}^{n} \left[X_i^2\right] - n\bar{X}^2}$$

Relationship between F-test and t-test

$$
\begin{aligned}
f &= \left(\frac{RSS_r - RSS_u}{RSS_u}\right)\left(\frac{n-k}{q}\right) \\
&= \left(\frac{\sum_{i=1}^{n}\left[X_i^2\right] - \left(\sum_{i=1}^{n}\left[X_i^2\right] - n\bar{X}^2\right)}{(n-1)\bar{S}^2}\right)\left(\frac{n-1}{1}\right) \\
&= \left(\frac{n\bar{X}^2}{(n-1)\bar{S}^2}\right)\left(\frac{n-1}{1}\right) = \frac{\bar{X}^2}{\left(\frac{\bar{S}^2}{n}\right)} \\
&= \left(\frac{\hat{\mu}_X - 0}{\sqrt{\frac{\bar{S}^2}{n}}}\right)^2 = \left(\frac{\hat{\mu}_X - 0}{se\left(\hat{\mu}_X\right)}\right)^2 = \left(t_{\hat{\mu}_X}\right)^2
\end{aligned}
$$

So, the F-statistic is *exactly* equal to the square of the t-statistic for the single hypothesis test with $H_0 : \mu_X = 0$ and $H_a : \mu_X \neq 0$ (in fact, the same turns out to be true for *any* single hypothesis test in regression inference).

Distribution of F-Statistic

- Consider the F-Statistic:

$$f = \frac{\left(\frac{RSS_r - RSS_u}{q}\right)}{\left(\frac{RSS_u}{n-k}\right)} = \frac{\left(\frac{RSS_r - RSS_u}{\sigma_X^2}\right)\left(\frac{1}{q}\right)}{\left(\frac{RSS_u}{\sigma_X^2}\right)\left(\frac{1}{n-k}\right)}$$

- If the X_is are *precisely* normally distributed then $\frac{RSS_r - RSS_u}{\sigma_X^2} \stackrel{d}{=} \chi^2(q)$ and $\frac{RSS_u}{\sigma_X^2} \stackrel{d}{=} \chi^2(n-k)$.

- Hence the F-statistic then has a *precise* $F(q, n-k)$ distribution:

$$f \stackrel{d}{=} F(q, n-k)$$

Distribution of F-Statistic

- However, we usually *cannot* assume that the X_is are normally distributed, so what is *much* more important for our purposes is the *asymptotic* distribution of the F-statistic.
- Firstly, by the Law of Large Numbers:

$$\frac{RSS_u}{n-k} \xrightarrow{p} \sigma_X^2 \Longrightarrow \left(\frac{RSS_u}{\sigma_X^2}\right)\left(\frac{1}{n-k}\right) \xrightarrow{p} 1$$

- This then means that, by the Central Limit Theorem:

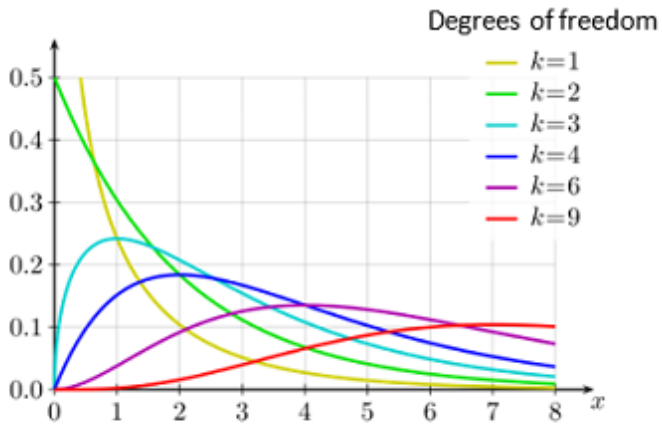
$$q \times f \xrightarrow{d} \chi^2(q)$$

- So the **crucial take-away point** here is that when you run an F-test with degrees of freedom $(q, n-k)$ on the computer (or are given the information to calculate the F-statistic and run the test), at high sample size then this is just equivalent to a chi-square test with q degrees of freedom.

Relationship between Chi-Square and F Distributions

Critical Values of Chi Square

df	0.05	0.01	df	0.05	0.01
1	3.84	6.64	16	26.30	32.00
2	5.99	9.21	17	27.59	33.41
3	7.82	11.34	18	28.87	34.80
4	9.49	13.28	19	30.14	36.19
5	11.07	15.09	20	31.41	37.57
6	12.59	16.81	21	32.67	38.93
7	14.07	18.48	22	33.92	40.29
8	15.51	20.09	23	35.17	41.64
9	16.92	21.67	24	36.42	42.98
10	18.31	23.21	25	37.65	44.31
11	19.68	24.72	26	38.88	45.64
12	21.03	26.22	27	40.11	46.96
13	22.36	27.69	28	41.34	48.28
14	23.68	29.14	29	42.56	49.59
15	25.00	30.58	30	43.77	50.89



Source: https://en.wikipedia.org/wiki/Chi-squared_distribution

Distribution of F (.05 Level of Significance)

	df_B								
df_W	1	2	3	4	5	6	7	8	12
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.18
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.16
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.13
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.12
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.00
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.92
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.88
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.85
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.83
∞	3.84	3.00	2.61	2.37	2.22	2.10	2.01	1.94	1.75
	1	2	3	4	5	6	7	8	12
df_W	df_B								

Note: Critical value of $\chi^2(k)$ is equal to k multiplied by critical value of $F(k, \infty)$, where $k = df = df_B$.

Testing Multiple Population Parameters

- Suppose we want to test the multiple hypothesis that average hourly earnings in *both* subgroups are equal to some specific value. Then we set $H_0 : \mu_m = \mu_f = \mu_0$ and $H_a : \mu_f \neq \mu_0 \vee \mu_m \neq \mu_0$.

- Under the null hypothesis H_0:

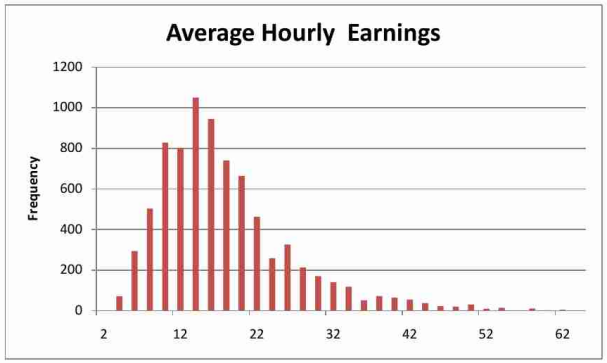
$$t_{\hat{\mu}_m} = \frac{\hat{\mu}_m - \mu_0}{se\left(\hat{\mu}_m\right)} = \frac{17.7726 - \mu_0}{\left(\frac{9.3036}{\sqrt{4673}}\right)} \sim N(0,1)$$

$$t_{\hat{\mu}_f} = \frac{\hat{\mu}_f - \mu_0}{se\left(\hat{\mu}_f\right)} = \frac{15.3586 - \mu_0}{\left(\frac{7.7100}{\sqrt{3313}}\right)} \sim N(0,1)$$

- So the $\chi^2(2)$ test statistic will then be:

$$W = (t_{\hat{\mu}_m})^2 + (t_{\hat{\mu}_f})^2 \sim \chi^2(2)$$

Testing for Normality (Jarque-Bera test)



Sample Statistics:

Mean	16.77
Median	14.90
Standard Deviation	8.76
Skewness	1.41
Excess Kurtosis	2.65
n	7986

Using the sample values of the skewness and kurtosis, it is possible to *test* whether the data are a random sample from the Normal distribution. If they are, the sample skewness $\hat{\kappa}_3$ and excess kurtosis $\hat{\kappa}_4$ should be close to zero.

The test is based on the results that, under the null hypothesis of normality:

$$W_s = \frac{n\hat{\kappa}_3^2}{6} \sim \chi^2(1); \quad W_k = \frac{n\hat{\kappa}_4^2}{24} \sim \chi^2(1); \quad W_n = W_s + W_k \sim \chi^2(2)$$

The 95% critical value for the $\chi^2(2)$ distribution is 5.99. For the earnings data above, evaluating the test statistic gives $W_n = 4996.86$.

Note: The formulas for sample skewness and sample excess kurtosis (unbiased estimators of population parameters) are quite complex due to degrees of freedom adjustment (statistical software packages calculate them for you anyway) but it is also possible to use $\hat{\kappa}_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\bar{X})^3\right]}{\left(\frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\bar{X})^2\right]\right)^{\frac{3}{2}}}$ as a consistent estimator for population skewness and $\hat{\kappa}_4 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\bar{X})^4\right]\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}\left[(X_i-\bar{X})^2\right]\right)^2} - 3$ as a consistent estimator for population excess kurtosis.