STAT 135 Lab 7
Distributions derived from the normal distribution, and comparing independent samples.

Rebecca Barter

March 16, 2015

The χ^2 distribution

The χ^2 distribution

We have seen several instances of test statistics which follow a χ^2 distribution, for example, the generalized likelihood ratio, Λ

$$-2\log(\Lambda) \sim \chi^2_{df}$$

and the Pearson χ^2 goodness-of-fit test statistic

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-1-dim(\theta)}$$

It turns out that the χ^2 distribution is very related to the $N(0,1)$ distribution.

The χ^2 distribution

The χ^2 distribution can be generated from the sum of squared normals:

$$\text{If } Z \sim N(0,1), \text{ then } Z^2 \sim \chi^2_1$$

and more generally, if $Z_i \sim N(0,1)$, then

$$\boxed{Z_1^2 + \ldots + Z_n^2 \sim \chi^2_n}$$

from which it is easy to conclude that if $U \sim \chi^2_n$ and $V \sim \chi^2_m$, then

$$U + V \sim \chi^2_{m+n}$$

that is, to get the distribution of the sum of χ^2 random variables, just add the degrees of freedom.

The χ^2 distribution

The χ^2_n distribution is also related to the gamma distribution:

$$U \sim \chi^2_n \iff U \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

so we can write the density of the χ^2_n distribution as

$$f(x) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad x \geq 0$$

The t distribution

The t distribution

If $Z \sim N(0,1)$ is independent of $U \sim \chi^2_n$, then

$$\frac{Z}{\sqrt{\frac{U}{n}}} \sim t_n$$

The t_n distribution has density

$$f_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

The t distribution

Note that if $X_i \sim N(\mu, \sigma^2)$, then

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

However, if we don't know σ, we can estimate it by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

and our distribution becomes

$$\frac{\bar{X}_n - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

The F-distribution

The F-distribution

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then

$$F = \frac{U/m}{V/n} \sim F_{m,n}$$

The $F_{m,n}$ distribution has density

$$f(w) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n} w\right)^{-(m+n)/2}$$

Moment generating functions

Moment generating functions

The moment generating function (MGF) for a random variable X is given by

$$M(t) = E\left(e^{tX}\right)$$

and the MGF of a random variable uniquely determines the corresponding distribution.

Moment generating functions

The MGF of X is defined by

$$M(t) = E\left(e^{tX}\right)$$

A useful formula:

- The kth derivative of the MGF of X, evaluated at zero, is equal to the kth moment of X:

$$E\left(X^k\right) = \left.\frac{d^k M}{dt^k}\right|_{t=0}$$

Moment generating functions

Some nice properties of the moment generating function (show them yourself!):

- If X has MGF $M_X(t)$, and $Y = a + bX$, then Y has MGF

$$M_Y(t) = e^{at} M_X(bt)$$

- If X and Y are *independent* with MGF's M_X and M_Y and $Z = X + Y$, then

$$M_Z(t) = M_X(t) M_Y(t)$$

Exercise

Exercise: moment generating functions

For each of the following distributions, calculate the moment generating function, and the first two moments.

1. $X \sim Poisson(\lambda)$
2. $Z = X + Y$, where $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ are independent
3. $Z \sim N(0, 1)$

Comparing Independent Samples **from two normal populations with common but unknown variance**

- t-test

Comparing independent samples

So far, we have been working towards making inferences about parameters from a single population.

For example, we have looked at conducting hypothesis tests for the (unknown) population mean μ, such as

$$H_0 : \mu = 0$$

$$H_1 : \mu > 0$$

Given a sample $X_1, ..., X_n$ from our population of interest, we would look at the sample mean, $\bar{X}_n$, to see if we have enough evidence against H_0 in favor of H_1 (for example, by calculating a p-value)

Comparing independent samples

What if we were instead interested in *comparing* parameters from two different populations?

Suppose that:

- the first population has (unknown) mean μ_1, and
- the second population has (unknown) mean μ_2.

Then we might test the hypothesis:

$$H_0 : \mu_1 = \mu_2$$

against

$$H_1 : \mu_1 > \mu_2 \quad \text{or} \quad H_1 : \mu_1 < \mu_2 \quad \text{or} \quad H_1 : \mu_1 \neq \mu_2$$

Comparing independent samples from two normal populations with common but unknown variance

Suppose that we have observed samples:

- $X_1 = x_1, ..., X_n = x_n$ from pop 1 with unknown variance σ_X^2
- $Y_1 = y_1, ..., Y_m = y_m$ from pop 2 with unknown variance σ_Y^2
- Assume common variance: $\sigma_X^2 = \sigma_Y^2$

We might want to test

$$H_0 : \mu_1 = \mu_2$$

against

$$H_1 : \mu_1 > \mu_2$$

What test statistic might we look at to see if we have evidence against H_0 in favor of H_1?

Comparing independent samples from two normal populations with common but unknown variance

- If H_0 was true, we might expect $\bar{X}_n - \bar{Y}_m \approx 0$
- if H_1 was true, we might expect $\bar{X}_n - \bar{Y}_m > 0$.

Note that our p-value (the probability, assuming H_0 true, of seeing something at least as extreme as what we have observed) would be given by

$$P_{H_0}(\bar{X} - \bar{Y} > \bar{x} - \bar{y})$$

To calculate this probability, we need to identify the distribution of $\bar{X} - \bar{Y}$.

Comparing independent samples from two normal populations with common but unknown variance

It turns out that

$$\boxed{t = \frac{(\bar{X}_n - \bar{Y}_m)}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}}$$

where the denominator is an estimate of the standard deviation of $\bar{X} - \bar{Y}$ (since the true variance is unknown), and

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

is the pooled variance of $X_1, ..., X_n$ and $Y_1, ..., Y_m$

Comparing independent samples from two normal populations with common but unknown variance

To test

$$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \mu_1 > \mu_2$$

Our p-value can be calculated by

$$P_{H_0}(\bar{X} - \bar{Y} > \bar{x} - \bar{y}) = P\left(\frac{(\bar{X} - \bar{Y})}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} > \frac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}\right)$$

$$= P\left(t_{n+m-2} > \frac{(\bar{x} - \bar{y})}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}\right)$$

This test is called a two-sample t-test

Exercise

Exercise: comparing independent samples from two normal populations with common but unknown variance

Suppose I am interested in comparing the average delivery time for two pizza companies. To test this hypothesis, I ordered 7 pizzas from pizza company A and recorded the delivery times:

$$(20.4\ ,\ 24.2\ ,\ 15.4\ ,\ 21.4\ ,\ 20.2\ ,\ 18.5\ ,\ 21.5)$$

and 5 pizzas from pizza company B:

$$(20.2\ ,\ 16.9\ ,\ 18.5\ ,\ 17.3\ ,\ 20.5)$$

I know that the delivery times follow normal distributions, but I don't know the mean or variance.
Do I have enough evidence to conclude that the average delivery times for each company are different?

A nonparametric test for comparing Independent Samples **from two arbitrary populations**

- Mann-Whitney test

Comparing Independent Samples from two arbitrary populations

Suppose that we are interested in comparing two populations, but we don't have any information about the distribution of our two populations.

We observe data

- $X_1 = x_1, ..., X_n = x_n$ IID with unknown cts distribution F
- $Y_1 = y_1, ..., Y_m = y_m$ IID with unknown cts distribution G

Suppose that we want to test the null hypothesis

$$H_0 : F = G$$

We could test H_0 using a nonparametric test (a test that makes no distributional assumptions on the data) based on ranks.

Comparing independent samples from two arbitrary populations

Let $\mathbf{R_i}$ **be the rank of** $\mathbf{X_i}$ (when the ranks are taken over all of the X_i's and Y_i's combined).

What is the expected value of $\sum_{i=1}^{n} R_i$ under H_0?

First note that since under H_0 all ranks are equally likely,

$$P(R_1 = k) = \frac{1}{n+m}$$

Thus

$$\begin{aligned} E(R_1) &= 1P(R_1 = 1) + 2P(R_1 = 2) + \ldots + (n+m)P(R_1 = n+m) \\ &= \frac{1}{n+m}\left(1 + 2 + \ldots + (n+m)\right) \\ &= \frac{1}{n+m}\left(\frac{(n+m)(n+m+1)}{2}\right) \\ &= \frac{n+m+1}{2} \end{aligned}$$

Comparing independent samples from two arbitrary populations

We just showed that

$$E(R_1) = \frac{n+m+1}{2}$$

So that if $H_0 : F = G$ is true, we have

$$E\left(\sum_{i=1}^{n} R_i\right) = \frac{n(n+m+1)}{2}$$

Thus a value of $\sum_{i=1}^{n} R_i$ that is significantly larger or smaller than $\frac{n(n+m+1)}{2}$ would provide evidence against $H_0 : F = G$.

Example: Comparing independent samples from two arbitrary populations

Suppose, for example, that we had

$$X = (6.2, 3.7, 4.7, 1.3)$$

$$Y = (1.2, 0.8, 1.4, 2.5, 1.1)$$

Then our ranks are

$$rank(X) = (9, 7, 8, 4)$$

$$rank(Y) = (3, 1, 5, 6, 2)$$

and we have that the sum of the ranks of the X_i's and Y_i's are

$$\text{sum of } rank(X) \ := \sum_{i=1}^{n} R_i = 28$$

$$\text{sum of } rank(Y) = \sum_{i=1}^{m} R_i' = 17$$

Comparing independent samples from two arbitrary populations

In our example, we saw that the sums of the ranks of the X_i's was larger than the sums of the ranks of the Y_i's.

But how can we tell if this difference is enough to conclude that $H_0 : F = G$ is false (that the two samples came from different distributions)?

Comparing independent samples from two arbitrary populations

Suppose that

- the rank of X_i is R_i
- the rank of Y_i is R'_i

Note that under H_0, if $n = m$, we would expect that

$$\sum_{i=1}^{n} R_i = \sum_{i=1}^{m} R'_i$$

Recall that

$$E\left(\sum_{i=1}^{n} R_i\right) = \frac{n(n+m+1)}{2}$$

similarly we should have that the expected sum of the R'_i's is

$$E\left(\sum_{i=1}^{m} R'_i\right) = \frac{m(n+m+1)}{2}$$

Comparing independent samples from two arbitrary populations

Thus if H_0 were true, we would expect that

$$\sum_{i=1}^{n} R_i + \sum_{i=1}^{m} R'_i \approx n_1(n+m+1)$$

where n_1 is the sample size of the smaller population:
$n_1 = \min(n, m)$.
Based on this idea, we can use the **Mann-Whitney test** to test H_0.

Comparing independent samples from two arbitrary populations

The Mann-Whitney test can be conducted as follows

1. Concatenate the X_i and Y_j into a single vector Z
2. Let n_1 be the sample size of the smaller sample
3. Compute $R =$ sum of the ranks of the smaller sample in Z
4. Compute $R' = n_1(m + n + 1) - R$
5. Compute $R^* = \min(R, R')$
6. Compare the value of R' to critical values in a table: if the value is less than or equal to the tabulated value, reject the null that $F = G$

TABLE 8 Critical Values of Smaller Rank Sum for the Wilcoxon Mann-Whitney Test

n_2	α for Two-Sided Test	α for One-Sided Test	n_1 (Smaller Sample) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.20	.10		3	7																	
	.10	.05			6																	
	.05	.025																				
	.01	.005																				
4	.20	.10		3	7	13																
	.10	.05			6	11																
	.05	.025				10																
	.01	.005																				
5	.20	.10		4	8	14	20															
	.10	.05		3	7	12	19															
	.05	.025			6	11	17															
	.01	.005					15															
6	.20	.10		4	9	15	22	30														
	.10	.05		3	8	13	20	28														
	.05	.025			7	12	18	26														
	.01	.005				10	16	23														
7	.20	.10		4	10	16	23	32	41													
	.10	.05		3	8	14	21	29	39													
	.05	.025			7	13	20	27	36													
	.01	.005				10	16	24	32													
8	.20	.10		5	11	17	25	34	44	55												
	.10	.05		4	9	15	23	31	41	51												
	.05	.025		3	8	14	21	29	38	49												
	.01	.005				11	17	25	34	43												
9	.20	.10	1	5	11	19	27	36	46	58	70											
	.10	.05		4	*10	16	24	33	43	54	66											
	.05	.025		3	8	14	22	31	40	51	62											
	.01	.005			6	11	18	26	35	45	56											

(continued)

TABLE 8 Critical Values of Smaller Rank Sum for the Wilcoxon Mann-Whitney Test (Continued)

n_2	α for Two-Sided Test	α for One-Sided Test	n_1 (Smaller Sample) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
10	.20	.10	1	6	12	20	28	38	49	60	73	87										
	.10	.05		4	10	17	26	35	45	56	69	82										
	.05	.025		3	9	15	23	32	42	53	65	78										
	.01	.005			6	12	19	27	37	47	58	71										
11	.20	.10	1	6	13	21	30	40	51	63	76	91	106									
	.10	.05		4	11	18	27	37	47	59	72	86	100									
	.05	.025		3	9	16	24	34	44	55	68	81	96									
	.01	.005			6	12	20	28	38	49	61	73	87									
12	.20	.10	1	7	14	22	32	42	54	66	80	94	110	127								
	.10	.05		5	11	19	28	38	49	62	75	89	104	120								
	.05	.025		4	10	17	26	35	46	58	71	84	99	115								
	.01	.005			7	13	21	30	40	51	63	76	90	105								
13	.20	.10	1	7	15	23	33	44	56	69	83	98	114	131	149							
	.10	.05		5	12	20	30	40	52	64	78	92	108	125	142							
	.05	.025		4	10	18	27	37	48	60	73	88	103	119	136							
	.01	.005			7	*13	22	31	41	53	65	79	93	109	125							
14	.20	.10	1	*8	16	25	35	46	59	72	86	102	118	136	154	174						
	.10	.05		*6	13	21	31	42	54	67	81	96	112	129	147	166						
	.05	.025		4	11	19	28	38	50	62	76	91	106	123	141	160						
	.01	.005			7	14	22	32	43	54	67	81	96	112	129	147						
15	.20	.10	1	8	16	26	37	48	61	75	90	106	123	141	159	179	200					
	.10	.05		6	13	22	33	44	56	69	84	99	116	133	152	171	192					
	.05	.025		4	11	20	29	40	52	65	79	94	110	127	145	164	184					
	.01	.005			8	15	23	33	44	56	69	84	99	115	133	151	171					
16	.20	.10	1	8	17	27	38	50	64	78	93	109	127	145	165	185	206	229				
	.10	.05		6	14	24	34	46	58	72	87	103	120	138	156	176	197	219				
	.05	.025		4	12	21	30	42	54	67	82	97	113	131	150	169	190	211				
	.01	.005			8	15	24	34	46	58	72	86	102	119	136	155	175	196				

Example: Comparing independent samples from two arbitrary populations

Back to our example, where

$$X = (6.2, 3.7, 4.7, 1.3) \quad rank(X) = (9, 7, 8, 4)$$

$$Y = (1.2, 0.8, 1.4, 2.5, 1.1) \quad rank(Y) = (3, 1, 5, 6, 2)$$

Here $n_1 = 4$.
The smaller sample is X and so the sum of the ranks of X are:

$$R = \text{ sum of } rank(X) := \sum_{i=1}^{n} R_i = 28$$

Next, we compute

$$R' = n_1(m + n + 1) - R = 4 \times (5 + 4 + 1) - 28 = 12$$

Example: Comparing independent samples from two arbitrary populations

$$R = \text{ sum of } rank(X) \ := \sum_{i=1}^{n} R_i = 28$$

$$R' = n_1(m + n + 1) - R = 4 \times (5 + 4 + 1) - 28 = 12$$

so that

$$R^* = \min(R, R') = 12$$

The critical value in the table for a two-sided test, with $\alpha = 0.05$ $(n_1 = 4, n_2 = 5)$ is 11. Thus, our value is not less than or equal to the critical value, so we fail to reject H_0.

Exercise

Exercise: Rice Chapter 11, Exercise 21

A study was done to compare the performances of engine bearings made of different compounds. Ten bearings of each type were tested. The times until failure (in units of millions of cycles) for each type are given below.

Type I	3.03	5.53	5.60	9.30	9.92
	12.51	12.95	15.21	16.04	16.84
Type II	3.19	4.26	4.47	4.53	4.67
	4.69	12.78	6.79	9.37	12.75

1. Use normal theory to test the hypothesis that there is no difference between the two types of bearings.
2. Test the same hypothesis using a nonparametric method.
3. Which of the methods – that of part (a) or that of part (b) – do you think is better in this case?