Stats 135, Fall 2019

Lecture 22, Wednesday, 10/23/2019

1 Review

Last time, we looked into a 2-sample t-test and we compared means of independent normal populations with different variances.

2 Today

Today we'll remove the assumption that they are independent, but we will assume they are paired (such as the same person, before and after something).
We'll look at the **paired** 2 sample t-test (paired and unpaired test in R). Then we'll jump back to §9.5 to consider the multinomial distribution and the chi Square test.

3 §11.3.1: Paired t-test

We wish to estimate $\mu_X - \mu_Y$ but in paired studies, our samples are no longer independent. Let $X_1, \ldots, X_n$ be iid $N(\mu_X, \sigma_X^2)$ and $Y_1, \ldots, Y_n$ iid $N(\mu_Y, \sigma_Y^2)$, and we'll assume their variances are unknown. Then we have

$$\sigma_{XY} = \text{Cov}(X_i, Y_i) \neq 0.$$

Let $D_i := X_i - Y_i$ so that

$$\begin{aligned} \mathbb{E}(D_i) &= \mu_X - \mu_Y \\ \text{Var}(D_i) &= \text{Var}(X_i - Y_i) = \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}, \end{aligned}$$

where we have the negative in the covariance term because $D_i = X_i - Y_i$. We basically have n individuals, so the average of their difference is approximately normal with:

$$\overline{D} \approx N\left(\mu_X - \mu_Y, \frac{1}{n}(\sigma_X^2 + \sigma^2 - 2\sigma_{XY})\right).$$

The only new idea here is that if X, Y are positively correlated, then

$$\sigma_{XY} > 0 \text{ and } \text{Var}(\overline{D}) < \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n},$$

which is $\text{Var}(\overline{X} - \overline{Y})$ when X, Y independent. Then we define efficiency as the ratio

$$\text{efficiency} := \frac{s_{\text{paired}}}{s_{\text{unpaired}}} < 1.$$

This results in a larger t.s., so we are more likely to reject the null for paired data. That is,

$$t = \frac{\overline{D} - \mu_D}{s_{\overline{D}}} \sim t_{n-1}.$$

If we do a hypothesis tails, we would design:

$$\begin{aligned} H_0 &: \mu_D = 0 \\ H_1 &: \mu_D \neq 0, \end{aligned}$$

which has rejection region $|\overline{D}| > t_{n-1}\left(\frac{\alpha}{2}\right) s_{\overline{D}}$.

4 §9.5: The Multinomial Distribution

This generalizes the binomial distribution. The multinomial random variable is a sum of n independent m outcome cells ($m = 2$ for binomial distribution). Let

$$p_1, p_2, \ldots, p_m$$

be the probabilities of outcome $1, 2, \ldots, m$ respectively, where $p_1 + \cdots + p_m = 1$.
We may ask, what is the chance of getting X_1 outcome 1, X_2 outcome 2, and so on until X_m outcome m? Then $n = x_1 + \cdots + x_m$. Define the generalization of 'choose' as :

$$\binom{n}{x_1, x_2, \ldots, x_m} := \frac{n!}{x_1! \cdots x_m!},$$

so that our desired probability is

$$\binom{n}{x_1, x_2, \ldots, x_m} p_1^{x_1} \cdots p_m^{x_m},$$

which we know as the multinomial formula. Now if we know $p_1, \ldots, p_m$, then the multinomial distribution is **completely specified**. In other words, we can calculate the probability of ANY outcome.
We draw a picture of the cells and compare what we observe to what we expect for each cell. Notice that for the observer, we have $\sum_{i=1}^{n} X_i = n$, and for the expected, $\sum_{i=1}^{m} p_i = 1$. Then we have

$$X_i \sim \text{Binomial}(n, p_i) \approx N(np_i, np_iq_i).$$

This brings us to the **chi Square statistic**, taking the observed minus expected counts over the expected:

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{n\to\infty} \chi^2_{m-1}.$$

When we have 2 outcomes (binomial), we will show this is χ^2_1. This gives

$$\sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2},$$

and a bit of algebra yields

$$\frac{(X_1 - np_1)^2}{np_1q_1} = \frac{(X_1 - \mathbb{E}(X_1))}{\text{Var}(X_1)} = z^2 \sim \chi^2_1.$$

Now more generally, we have:

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \overset{\text{algebra}}{=} \sum_{i=1}^{m-1} \frac{(O_i - E_i)^2}{E_iq_i},$$

which involves some linear algebra trickery or magic (this result is very nontrivial). Then for large n, we have

$$\sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{m-1} z_i^2 \sim \chi^2_{m-1}.$$

Now Lucas wants to give a bit of insight. We enter discussion through the example of fitting the emissions of alpha particles to the Poisson distribution. We take the null hypothesis to be that our observed counts come from the multinomial distribution $(1207, p_1, \ldots, p_{16})$. The alternative H_A is that the observed counts come from some other distribution.
Now we calculate our χ^2 statistic and we will add over all the rows:

$$\sum_{\text{all rows}} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{16-1-1}$$

where $16-1$ is the dimension of the sample space, and the 1 is the dimension of the null space (Poisson has 1 parameter).