Section 3: Chi-Squared Distribution, t-Tests, Confidence Intervals

TF: Reagan Rose (rrose@g.harvard.edu)

September 20, 2015

Chi-Squared Distribution

The **chi-squared distribution** is fundamental to all of statistics, but in particular *linear models*. This distribution arises naturally in trying to estimate the variance of normal random variables. That is, if we assume that our data is generated by a normal distribution, then our estimate for variance will generally follow a chi-squared distribution.

Properties of the chi-squared distribution

- **Sum of squared normals**: If $Z_1, \ldots, Z_n \sim^{i.i.d.} N(0,1)$, then $Y = Z_1^2 + \ldots + Z_n^2 \sim \chi_n^2$, where χ_n^2 is the chi-squared distribution with n degrees of freedom.
- **Special case of gamma:** The chi-squared distribution with n degrees of freedom is equivalent to a gamma distribution with parameters $(n/2, 1/2)$. That is,

$$\chi_n^2 = Gamma(n/2, 1/2).$$

Now as for how the chi-squared distribution arises in data analysis, the fundamental result is that if $X_1, \ldots, X_n \sim^{iid} N(\mu, \sigma^2)$, and we define the sample variance as

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

then

$$s^2 \sim \frac{\sigma^2}{n-1}\chi_{n-1}^2.$$

Main message: Just as our estimate of the mean of the normal distribution is distributed normal $\bar{X} \sim N(\mu, \sigma^2/n)$, out estimate of variance is distributed χ_{n-1}^2. We'll use this fact to construct new test statistics, run hypotehesis tests, create confidence intervals, and more!

t Distribution and the t-Test

Another reason the chi-squared distribution is so important is that is used to "build" the **Student's t distribution** (later you will see that the F distribution is also built upon the chi-squared).

Properties of the t-distribution

- **Ratio of normal to chi-squared**: If $Z \sim N(0,1)$ and $V \sim \chi^2_v$ with Z and V independent, then:
$$T = \frac{Z}{\sqrt{V/v}} \sim t_v$$
which is the t distribution with v degrees of freedom.
- **Zero-centered, symmetric**: By definition, the t distribution is centered at zero and is symmetric about zero.

Problem 1 (d) in your homework asks you to derive the sampling distribution of the statistic

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Hint: It should be a t distribution! This is why we use the t statistic and t distribution for comparing means of observed data. This should also show you why we use the t distribution (and t-statistics) sometimes rather than using the normal distribution (and z statistics) for hypothesis testing and confidence intervals.

Main message: When we know the true variance of our data, our test statistic $z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is *normally distributed*. But when we don't know the true variance (which is almost always the case), our test statistic $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ is *t-distributed*. The $n \geq 30$ rule of when to use the z versus t distribution comes from the fact that

$$lim_{n\to\infty} t_{n-1} = N(0,1).$$

This leads us directly to the **one-sample t-test**, which goes as follows:

1. Under the null hypothesis $H_0 : \mu = \mu_0$, we derive the sampling distribution of t. This tells us the values that t can possibly take on and how likely each value is (assuming the null hypothesis is true).
2. We then observe our data and calculate t_{obs}. In order to determine whether this value is *reasonable* under the null hypothesis, we see where it lies on the sampling distribution.
3. If it is very unlikely ($p < \alpha$) that we would see t_{obs} (or more extreme) under the null hypothesis, then we reject H_0. If not, then we say we do not have enough evidence to reject it. (Note that the variance of the t distribution gets smaller as n gets bigger, so even if we observe the exact same t_{obs} for larger n, we may be able to reject! This is why saying "we do not have enough evidence to reject" makes sense.)

We can extend this to the **two-sample t-test**, which takes two forms:

- **unpaired (or independent) two-sample t-test:**

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2 | H_0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \approx min(n_1 - 1, n_2 - 1).$$

- **paired two-sample t-test:** This is used if we want to compare measurements *of the same subject*. These tests are powerful, but very rarely used. Formulas for these test will be discussed later.

Confidence Intervals

The goal of interval estimation is to provide an interval over parameter values (which we will denote as θ) that covers the true value θ_0 with some probability. So if we are interested iestimating the mean μ_0 of a normal distribution, a confidence interval will tell us what values of μ_0 are "reasonable" given our data. Or in other words, what values of μ_0 make our data seem probable? In this case, we would find the values of μ_0 such that

$$\alpha/2 \leq P\left(T \geq \frac{\bar{x} - \mu_0}{s/sqrtn}\right) \leq 1 - \alpha/2.$$

For t-distributed data, the structure of a $(1 - \alpha)\%$ confidence interval is

$$\bar{x} - t_{1-\alpha/2, n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2, n-1}\frac{s}{\sqrt{n}}.$$

Warning: tricky semantics: A confidence interval should **NOT** be interpreted as

Given a confidence interval, it will contain the true value θ with probability $1 - \alpha$.

Instead, what a confidence interval tells us is

If we draw a large number of samples and construct confidence intervals for each sample, then $1 - \alpha$ of these confidence intervals will contain the true value.

Does this sound confusing? It is. In fact, some people were so confused/unhappy with this concept that they formed a whole different branch of statistics to get around it! This is where the Frequentist/Bayesian split in statistics generally starts!

Exercises

Question 1. Understanding the chi-squared Derive the distribution of the sample variance for normal data. I.e. Prove that

$$s^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}.$$

Question 2. Two sample confidence interval Recall our two-sample t-statistic is

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2|H_0)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

We approximate the degrees of freedom as $df = min(n_1 - 1, n_2 - 1)$. Derive the appropriate $(1-\alpha) * 100\%$ confidence interval for $\mu_1 - \mu_2$.

Question 3. Simulating confidence intervals To understand the true meaning of a confidence interval, simulate 1,000 samples, each of size $n = 25$, from a standard normal distribution. For each of the 1,000 samples, calculate a confidence interval for the mean of the distribution. Try to plot the confidence intervals to see how they differ. How many of the 1,000 confidence intervals captured the true mean of the distribution, $\mu_0 = 0$?