χ^2, F, t and Degrees of Freedom

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

$\chi^2(N)$

- If $Z_1, Z_2, \ldots, Z_N \sim N(0,1)$ are iid, then
$$\sum Z_i^2 \sim \chi^2(N)$$
- Obvious application: $X_1, X_2, \ldots, X_N \sim N(\mu, \sigma^2)$
$$\left(\frac{X_1 - \mu}{\sigma}\right)^2 + \cdots + \left(\frac{X_N - \mu}{\sigma}\right)^2 \sim \chi^2(N)$$

Non-independent summands

- Consider $X_1, X_2, \dots, X_N \sim N(\mu, \sigma^2)$
- Compute $\left(\frac{X_1 - \bar{X}}{\sigma}\right), \left(\frac{X_2 - \bar{X}}{\sigma}\right), \dots, \left(\frac{X_N - \bar{X}}{\sigma}\right)$
- $\left(\frac{X_i - \bar{X}}{\sigma}\right) \sim N(0,1)$ but the variables are not indep.
  - Because if we know $\left(\frac{X_1 - \bar{X}}{\sigma}\right), \left(\frac{X_2 - \bar{X}}{\sigma}\right), \dots, \left(\frac{X_{N-1} - \bar{X}}{\sigma}\right)$, we can compute $\left(\frac{X_N - \bar{X}}{\sigma}\right)$ (**Problem: how?)**
- Degrees of Freedom (one view): the number of independent variables in the sum
  - $\left(\frac{X_1 - \bar{X}}{\sigma}\right)^2 + \cdots + \left(\frac{X_N - \bar{X}}{\sigma}\right)^2 \sim \chi^2(N-1)$

Degrees of Freedom

- Degrees of Freedom: another view
  (# of variables)-(# of constraints)
  - Variables: $\left(\frac{X_1-\bar{X}}{\sigma}\right), \left(\frac{X_2-\bar{X}}{\sigma}\right), \ldots, \left(\frac{X_N-\bar{X}}{\sigma}\right)$
  - Constraint: $\sum_{i=1}^{N}\left(\frac{X_i-\bar{X}}{\sigma}\right)$

- **Problem:** $W \sim \chi^2(k)$. What's $E(W)$?

Degrees of Freedom

- $E\left(\sum_{i=1}^{N}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2\right)=(N-1)$
- $E\left(\sum_{i=1}^{N}\left(\frac{X_i-\mu}{\sigma}\right)^2\right)=N$

- Makes sense: $|X_i-\bar{X}|$ is smaller than $|X_i-\mu|$

Residuals

- The sum of the squared standardized residuals is distributed with a χ^2 distribution
- Standardized residuals when we estimate the mean using the sample mean:
$$r_1 = \left(\frac{X_1 - \bar{X}}{\sigma}\right), r_2 = \left(\frac{X_2 - \bar{X}}{\sigma}\right), \ldots, r_{N=}\left(\frac{X_N - \bar{X}}{\sigma}\right)$$
- $\sum_{i=1}^{N} r_i^2 \sim \chi^2(N-1)$

ANOVA and χ^2

- SST (Total Sum of Squares)

  $$\sum_{ij}\left(X_{ij}-\bar{X}\right)^2: \;\; \frac{\sum_{ij}\left(X_{ij}-\bar{X}\right)^2}{\sigma^2} \sim \chi^2(Npoints-1)$$

  - Constraint: $\sum_{ij}\left(X_{ij}-\bar{X}\right)=0$
- RSS/SSE (Residual Sum of Squares/Error Sum of Squares)

  $$\sum_{ij}\left(X_{ij}-\bar{X}_i\right)^2: \;\; \frac{\sum_{ij}\left(X_{ij}-\bar{X}_i\right)^2}{\sigma^2} \sim \chi^2(Npoints-Ngroups)$$

  - Constraint: $\sum_j\left(X_{ij}-\bar{X}_i\right)=0$ for every j
- SSG/SSR (Group Sum of Squares/Regression Sum of Squares)

  `$\sum_i \; N_i(\bar{X}_i-\bar{X})^2=\sum_{ij}(\bar{X}_i-\bar{X})^2$

ANOVA and χ^2

- SST = SSE + SSG (exercise/book)
- $\frac{SST}{\sigma^2} \sim \chi^2(Npoints - 1)$, $\frac{SSR}{\sigma^2} \sim \chi^2(Npoints - Ngroups)$
- So $\frac{SSG}{\sigma^2} \sim \chi^2(Ngroups - 1)$

F distribution and ANOVA

- If $W_1 \sim \chi^2(k_1)$ and $W_2 \sim \chi^2(k_2)$, then
$$F = \frac{W_1/k_1}{W_2/k_2} \sim F(k_1, k2)$$
- This lets us get away with not actually knowing σ:
$$F = \frac{SSG}{\sigma^2(Ngroups - 1)} \Big/ \frac{SSE}{\sigma^2(Npoints - Ngroups)}$$

$$F \sim F(Ngroups - 1, Npoints - Ngroups)$$

t Distribution

- If $Z \sim N(0,1)$ and $W \sim \chi^2(k)$ and they're indep.,
$$T = \frac{Z}{\sqrt{W/k}} \sim t(k)$$
- $\frac{(N-1)s_X^2}{\sigma^2} = \sum \left(\frac{(X_i - \bar{X})}{\sigma} \right)^2 \sim \chi^2(N-1)$
- So that $\frac{(X_i - \mu)/\sigma}{\sqrt{s_X^2/\sigma^2}} = \frac{X_i - \mu}{s_X} \sim t(N-1)$