ANOVA - analysis of variance

Basic 1-way ANOVA.

k = # of samples

$i = 1, 2, \ldots, k$, for each i, $j = 1, 2, \ldots, n_i$

n_i = size of i^{th} sample

---

sample.

$i=1$ $\quad x_{11}, x_{12}, x_{13}, \ldots, x_{1n_1}$

$i=2$ $\quad x_{21}, x_{22}, \ldots, x_{2n_2}$

$\vdots \qquad \vdots$

$i=k$ $\quad x_{k1}, x_{k2}, \ldots, x_{kn_k}$

Assume: $X_{ij} \sim N(\mu, \sigma^2)$

$$\bar{X}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} = \text{sample mean for } i^{th} \text{ sample}$$

$$\bar{X}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{1} \sum_{j=1}^{n_i} X_{ij} = \text{overall sample mean}$$

$$N = \sum_{i=1}^{k} n_i$$

note.

$$\bar{X}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{n_i}\right)$$

$$\bar{X}_{\cdot\cdot} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

Recall: $x_i \sim N(\mu, \sigma^2) \Rightarrow z = \frac{x_i - \mu}{\sigma} \sim N(0,1)$

$z^2 \sim \chi^2(\nu = 1)$ (Thm 8.7 p. 241)

Also that. $\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma}\right)^2 \sim \chi^2(\nu = n-1)$ (Thm 8.4 p. 242)

Thus:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_{..})^2}{\sigma^2} \sim \chi^2(\nu = \textstyle\sum n_i - 1) \qquad N-1$$

$$\sum_{i=1}^{k} n_i \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{\sigma^2} \sim \chi^2(\nu = k-1)$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_{i.})^2}{\sigma^2} \sim \chi^2(\nu = \textstyle\sum n_i - k)$$

$\chi^2(\nu = n_i - 1)$

$\sum_{i=1}^{k}(n_i - 1) = \sum n_i - k = N - k$

We can also show that:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_{..})^2}{\sigma^2} = \sum_{i=1}^{k} n_i \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{\sigma^2} + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{(x_{ij} - \bar{x}_{i.})^2}{\sigma^2}$$

df: $(N-1) = (k-1) + (N-k)$

Also recall: $X_1 \sim \chi^2(\nu_1)$, $X_2 \sim \chi^2(\nu_2)$ $\Rightarrow \dfrac{X_1/\nu_1}{X_2/\nu_2} \sim F_{\nu_1,\nu_2}$ (3) Thm 8.14 p. 247

(ν_1 ↑ num. d.f., ν_2 ↑ denom. d.f.)

Thus: $$\frac{\sum_{i=1}^{k} n_i \left(\frac{\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}}{\sigma}\right)^2 \cdot \frac{1}{k-1}}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(\frac{X_{ij} - \bar{X}_{i\cdot}}{\sigma}\right)^2 \cdot \frac{1}{N-k}} \sim F_{k-1,\,N-k}$$

note σ cancels out...

Define: $$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{\cdot\cdot})^2$$ (sum of squares total)

$$SSTr = \sum_{i=1}^{k} n_i (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$ (sum of squares treatment)

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2$$ (sum of squared errors)

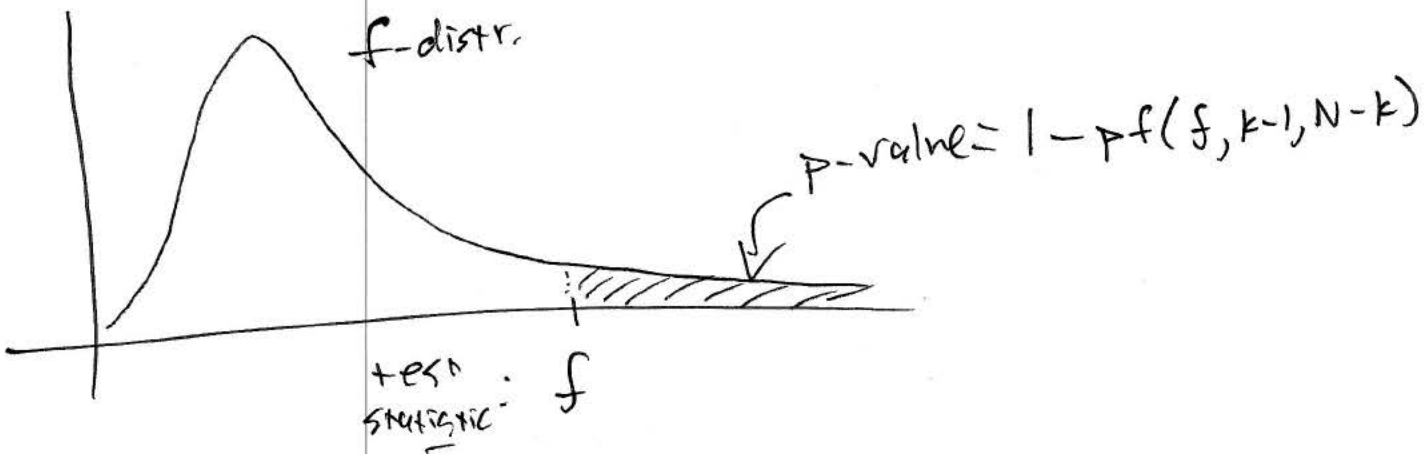
$$SST = SSTr + SSE$$

If we had a similar eq. for sum of sq's w/ lin. reg

Great: $$\frac{\frac{SSTr}{\sigma^2(k-1)}}{\frac{SSE}{\sigma^2(N-k)}} = \frac{SSTr/(k-1)}{SSE/(N-k)} \sim F_{k-1,N-k}$$

so what we do is we calculate a test statistic

$$f = \frac{\frac{SSTr}{k-1}}{\frac{SSE}{N-k}} \sim F_{k-1, N-k}$$



If test stat. f is too large, then we think our data is unusual.

The best explanation is that

$$X_{ij} \sim N(\mu, \sigma^2) \quad \text{is false.}$$

but we keep the assumption of Normality and σ_i^2

so at least one of the data sets must have a mean that isn't μ.

Explanation:

$\bar{X}_{..} = \hat{\mu}$
$\bar{X}_{i.} = \hat{\mu}$
} both estimators of μ

If we assume $X_{ij} \sim N(\mu_i, \sigma^2)$ $i = 1, 2, \ldots, k$

Then $\bar{X}_{..}$ estimates $\frac{1}{N}\sum_{i=1}^{k} n_i \mu_i = \bar{\mu}$

$\bar{X}_{i.}$ estimates μ_i

We aren't concerned w/ SSE being too large.

$(X_{ij} - \bar{X}_{i.})^2$ estimates $(X_{ij} - \mu_i)^2$

& we absolutely assume

$X_{ij} \sim N(\mu_i, \sigma^2)$

$(\bar{X}_{i.} - \bar{X}_{..})^2$ estimates $(\mu_i - \bar{\mu})^2$

If $\mu_i \neq \mu_j$ for at least one i, j pair,
then $\bar{\mu}$ & μ_i can be different
& SSTr might be large, even
for large samples. (since $\bar{X}_{i.}$ & $\bar{X}_{..}$ are estimating diff. things)

So we only worry if f is too large.

Notes:

- Test is more powerful

  i.e. lower type II error rates when sample size are identical, $n_i = n \quad i = 1, 2, \ldots, k$

- Test is fairly robust to data being non-normal or having unequal variances

  i.e. it's type I error rate will be close to α still

Rules of Thumb:

→ Keep sample sizes similar (say within a factor of 2x)

→ If sample variances differ by a factor of 4x or more, be wary of test result.

→ If data extremely skewed w/ outliers, be wary of test result.