χ^2 and t distribution

- U ~ χ^2_n and V ~ χ^2_m, then $U+V$ ~ χ^2_{m+n}
- Z ~ $N(0,1)$ and U ~ χ^2_n, then $\frac{Z}{\sqrt{\frac{U}{n}}}$ is t distribution with n degrees of freedom
- U ~ χ^2_n and V ~ χ^2_m, then $W = \frac{\frac{U}{m}}{\frac{V}{n}}$ is F distribution with m and n degrees of freedom
- $\sum_{i=1}^n (\frac{X_i-\mu}{\sigma})^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i-\mu)^2 \sim \chi^2_n$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
- $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Theory:

- $\overline{X}$ and $(X_1-\overline{X}, X_2-\overline{X}, \ldots, X_n-\overline{X})$ are independent
- $\overline{X}$ and S^2 are independently distributed

linear regression

- assumption: $y = x\beta + e$ where x is fixed, full rank matrix, e is homoscedastic random vector, with $e_1, \ldots e_n \overset{i.i.d}{\sim} N(0,\sigma^2)$
- $\hat{y} = x\hat{\beta}$, regression line, where $\hat{\beta} = (x'x)^{-1}x'y$
- $\sum_{i=1}^n (\frac{e_i}{\sigma})^2 \sim \chi^2_n$
  $\frac{RSS}{\sigma^2} = \frac{\sum_{i=1}^n \hat{e_i}^2}{\sigma^2} = \chi^2_{n-2}$
- $E(\hat{\beta}) = \beta, var(\hat{\beta}) = \sigma^2(x'x)^{-1}$

- $\hat{\beta}_0 \sim N(\beta_0, \sigma^2[\frac{1}{n} + \frac{\overline{x}^2}{nvar(x)}])$

- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{nvar(x)})$

- $cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2\overline{x}}{nvar(x)}$

- $E(\chi^2_{n-2}) = n - 2$
  $\Rightarrow E(\frac{RSS}{\sigma^2}) = n - 2$
  $\Rightarrow \hat{\sigma}^2 = \frac{RSS}{n-2}\ is\ an\ unbiased\ estimator\ of\ \sigma^2$

- $Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{nvar(x)}}}\ and\ U = \frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$
  $\Rightarrow t_{n-2} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_0 - \beta_0}{S_{\hat{\beta}_0}}$
  $S_{\hat{\beta}_1} = \sqrt{\frac{RSS}{n(n-2)var(x)}}$
  $S_{\hat{\beta}_0} = \sqrt{\frac{RSS}{n-2}(\frac{1}{n} + \frac{\overline{x}^2}{nvar(x)})}$

- $S_{\hat{y}_i} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{nvar(x)}}$

Comparing two independent samples

- if X, Y have the same variance σ^2, then $\overline{X} - \overline{Y} \sim N(\mu_x - \mu_y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$

- if σ^2 is not known, the pooled sample variance is an unbiased estimation of σ^2 :
  $s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$

- (two sample t test) $\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$

- Hypothesis testing: $H_0 : \mu_x - \mu_y = 0$
  $H_1 : \mu_x - \mu_y \neq 0$

- $t = \frac{\overline{X} - \overline{Y} - 0}{s_{\overline{X} - \overline{Y}}} \sim t_{n+m-2}$

- if X, Y have different variances, then estimator of Var is: $s_{\overline{X} - \overline{Y}} = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$
  $df = \frac{[(s_X^2/n) + (s_Y^2/m)]^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}$

- $H_0 : \mu_x - \mu_y = 0$ power $= P_1(d(x) = 1)$
  $H_1 : \mu_x - \mu_y = \Delta$ $\quad = P_1(\overline{X} - \overline{Y} > z(\alpha)\sigma\sqrt{\frac{2}{n}})$
  $= P_1(\frac{\overline{X} - \overline{Y} - \Delta}{\sigma\sqrt{\frac{2}{n}}} > \frac{z(\alpha)\sigma\sqrt{\frac{2}{n}} - \Delta}{\sigma\sqrt{\frac{2}{n}}})$
  $= 1 - \Phi(z(\alpha) - \frac{\Delta}{\sigma\sqrt{\frac{2}{n}}})$

one way ANOVA test

- parametric

- assumption: the data in each treatment group is independent, normal and with equal variance

- H_0 : the mean of each group is the same

  H_1 : at least one of the means is different

Mann Whitney test

- non parametric test version of unpaired t test

- don't assume normality of our data

- $H_0 : F = G$
  $H_1 : F \neq G$

- $\pi = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} I(x_i < y_j) = \frac{1}{mn} U_y$

- $U_y = T_y - \frac{m(m+1)}{2}$
  $U_y \sim N(\frac{mn}{2}, \frac{mn(m+n+1)}{12})$