Lecture 22

Topics: Finish hypothesis testing

Summary of the hypothesis testing

All tests are of the following form:

- Population parameter: θ
- Null hypothesis H_0: θ = some fixed number
- Estimate of θ based on samples from the population: $\hat{\theta}$
- Approximate $z = \dfrac{\hat{\theta} - E(\hat{\theta})}{sd(\hat{\theta})}$ assuming H_0 is true.
- Convert the (approximate) z-score to p-value by studying the sampling variability of z (or the approximation of z).
- if the p-value is really small (less than some cut off, say), i.e. if $|z|$ is large, then this is evidence against H_0.
  If the p-value is not small (greater than some cut off, say) then the data is consistent with H_0 (but H_0 could still be false).

List of approximate z-scores & their sampling variability for the Non-exact "quick guess method".

Population Parameter	estimate	approximate z-score	approximate sampling variability
(1) Population average μ (population s.d. σ known)	$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$	$z = \frac{\hat{\mu} - \mu}{(\sigma/\sqrt{n})}$	$z \approx N(0,1)$.
(2) Population average μ (population s.d. σ unknown)	$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$, $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}$	$z = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}$	$z \approx N(0,1)$
(3) Population proportion p	$\hat{p} = \frac{X_1 + \cdots + X_n}{n}$ where $X_i = \begin{cases} 0 \\ 1 \end{cases}$	$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	$z \approx N(0,1)$
(4) The difference btwn two population averages $\mu_x - \mu_y$	$\hat{\mu}_x - \hat{\mu}_y$ where $\hat{\mu}_x = \frac{X_1 + \cdots + X_n}{n}$, $\hat{\mu}_y = \frac{Y_1 + \cdots + Y_m}{m}$, $\hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu}_x)^2}$, $\hat{\sigma}_y = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \hat{\mu}_y)^2}$	$z = \frac{(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)}{\sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}}}$	$z \approx N(0,1)$.

These are all of the form $\frac{\hat{\theta} - E(\hat{\theta})}{sd(\hat{\theta})}$ where MF1 & MF2 are used to compute $E(\hat{\theta})$ & $sd(\hat{\theta})$.

Modifications and assumptions to make the "quick guess method" official

For (1): You need n large enough, usually $n > 30$, for the CLT to kick in so $Z \approx N(0,1)$ becomes a good approximation.

For (2): If you assume the population to be normally distributed Then Z is exactly a draw from $t^{(n-1)}$, i.e. a t-distribution with $n-1$ degrees of freedom.

Note: if $n > 30$ then $t^{(n-1)}$ is approx Normal so $Z \approx N(0,1)$ works well.

For (3): You need n large (typically larger than what is needed for (1)) for the CLT to imply $Z \approx N(0,1)$)

For (4): To get an exact pdf for Z, You need to assume both populations are Normal and $\sigma_x = \sigma_y$. Now replace

$$\sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}} \quad \text{with} \quad \hat{\sigma}_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

Where $\hat{\sigma}_p$ is the "pooled est of s.d. from both samples $X_1, \ldots, X_n, Y_1, \ldots, Y_m$" defined as

$$\hat{\sigma}_p := \sqrt{\frac{n-1}{n+m-2}\hat{\sigma}_x^2 + \frac{m-1}{n+m-2}\hat{\sigma}_y^2}.$$

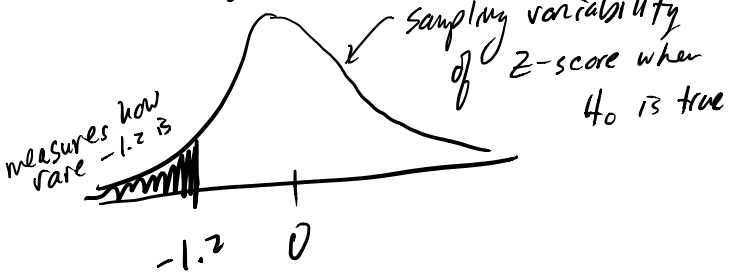
Now the "new" $Z = \dfrac{(\hat{\mu}_x - \hat{\mu}_y) - (\mu_x - \mu_y)}{\hat{\sigma}_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$

is exactly a draw from $t^{(n+m-2)}$.

All the p-values done in this class are "one-sided"

To convert a approximate Z-score to a p-value I will always use only one side of the pdf for Z:

e.g. if the approximate Z-score $= -1.2$ then the p-value is given by the shaded region below



e.g. if the approximate Z-score $= 2.5$ then the p-value is given by the shaded region below