# Stats 135, Fall 2019
## Lecture 17, Monday, 10/7/2019

Absent (out of state) - See attached written notes by Lucas for today's and Wednesday's lectures.

**CLASS ANNOUNCEMENTS:** Today marks the end of the materials that will be on the Midterm exam.

stat 135 lec17

Last time sec 9.4 GLRT

$H_0: \theta \in \omega_0$

$H_1: \theta \in \omega_1$

$\Omega = \omega_0 \cup \omega_1$

Fix $\alpha$ level of significance

We reject $H_0$ if $\Lambda = \dfrac{\max_{\theta \in \omega_0} \{lik(\theta)\}}{\max_{\theta \in \Omega} \{lik(\theta)\}} < c$

otherwise accept $H_0$.

Equivalently,

GLRT: reject $H_0$ if $-2\log\Lambda > K$ else accept $H_0$.

where

$-2\log\Lambda \sim \chi^2_{\dim\Omega - \dim\omega_0}$ and $K = \chi^2_{\dim\Omega - \dim\omega_0}(\alpha)$

Discussion Question (T, F explain)

The GLR, $\Lambda$, is always $\leq 1$ ?

True, the set in the denom includes the set in the numerator so the max of denom $\geq$ max of num.

End of material on midterm.

Sec 11.2 2 sample z, t test

1 sample z, t test.

## Sec 11.2 Comparing 2 indep samples

**Picture**

$\boxed{X \sim N(\mu_x, \sigma^2)}$ — drawn $n$ w/ replacement

$\boxed{Y \sim N(\mu_y, \sigma^2)}$ — drawn $m$ w/ replacement

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

since linear combinations of indep normal is normal and variances add.

If $\sigma^2$ is known

$100(1-\alpha)\%$ CI for $\mu_x - \mu_y$ is

$$\bar{X} - \bar{Y} \pm z\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

If $\sigma^2$ not known an unbiased estimate of $\sigma^2$ is the "pooled sample variance".

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Chk
$$E(S_p^2) = \frac{(n-1)\overset{=\sigma^2}{E(S_x^2)} + (m-1)\overset{=\sigma^2}{E(S_y^2)}}{n+m-2}$$

$$= \sigma^2 \checkmark$$

So $S_p^2$ is an unbiased estimator of $\sigma^2$.

We use this particular unbiased estimator because of the following theorem.

**Thm A** (2 sample t test)

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma^2)$

$Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma^2)$

The T.S.

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\underbrace{S_p \cdot \sqrt{\frac{1}{n} + \frac{1}{m}}}_{S_{\bar{X}-\bar{Y}}}} \sim t_{n+m-2}$$

Proof/ recall $\frac{Z}{\sqrt{U/n}} \sim t_n$ where $\left.\begin{matrix} Z \sim N(0,1) \\ U \sim \chi^2_n \end{matrix}\right\}$ indep.

and $\left.\begin{matrix} \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1} \\ \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{m-1} \end{matrix}\right\}$ indep since X and Y are indep.

$$\Rightarrow \frac{(n-1)S_X^2}{\sigma^2} + \frac{(m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

let $Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sigma \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1)$

$$U = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

then $\dfrac{Z}{\sqrt{U/_{n+m-2}}} \sim t_{n+m-2}$ by def$^n$.

since $Z$ and $U$ are independent

since $\bar{X}$ is indep of $S_X^2$ and $\bar{Y}$ is indep of $S_Y^2$ (see lec 8).

It remains to show

$$\frac{Z}{\sqrt{U/_{n+m-2}}} = \frac{\bar{X}-\bar{Y} - (\mu_X - \mu_Y)}{S_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}}$$

$$\frac{\bar{X}-\bar{Y} - (\mu_X - \mu_Y)}{S_p \cdot \sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{\bar{X}-\bar{Y} - (\mu_X - \mu_Y)}{\frac{1}{\sigma}\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}} \cdot \sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}$$

$$= \underbrace{\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}_{= Z} \cdot \underbrace{\frac{1}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{\sigma^2(m+n-2)}}}}_{= \frac{1}{\sqrt{U/_{m+n-2}}}} \quad \square$$

From thm A:

$$S_{\bar{X}-\bar{Y}} = S_p \sqrt{\frac{1}{n} + \frac{1}{m}}$$

<u>Cor A</u>

A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ is $\bar{X} - \bar{Y} \pm t_{m+n-2}\left(\frac{\alpha}{2}\right) \cdot S_{\bar{X}-\bar{Y}}$

Hypothesis testing for 2 sample problem using t-test.

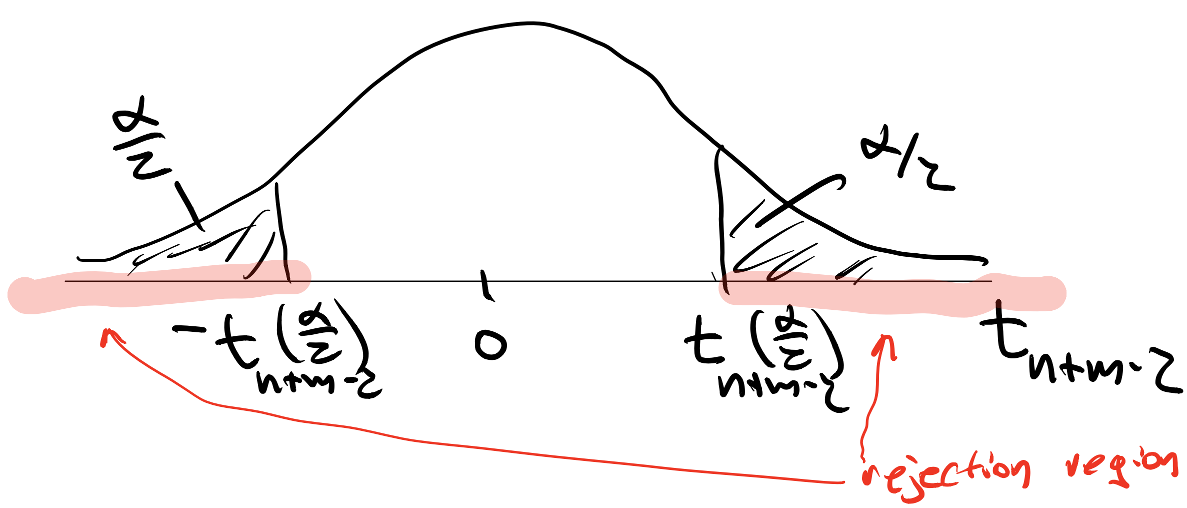
$H_0: \mu_X - \mu_Y = 0$
$H_1: \mu_X - \mu_Y \neq 0$ (2 sided alternative)

$\mu_X - \mu_Y > 0$
$\mu_X - \mu_Y < 0$ (1 sided alternatives)

For $\alpha$ level of significance
The T.S. used to make a decision whether to reject the null, assuming the null is true, is

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

Picture (2 sided alternative)




we reject $H_0$ if

$$|t| > t_{n+m-2}\left(\frac{\alpha}{2}\right) \quad \text{or}$$

equivalently

$$0 \notin (\bar{X} - \bar{Y}) \pm t_{n+m-2}\left(\frac{\alpha}{2}\right) S_{\bar{X}-\bar{Y}}$$

by duality of HT and CI.

Note if $n, m$ are large
the t-test is approximately a z-test.

Fact The t, z test is equivalent to the GLRT (proof in book p 426).

1 Sample Z test

This is a simple variation of a 2 sample t test where we have just 1 large sample.

**Problem** **(gender pay gap).** The average weekly earnings for a female social worker is \$670. Do men in the same positions have average weekly earnings that are higher than those for women? A random sample of $n = 40$ male social workers showed sample mean $\overline{X} = \$725$ and sample standard deviation $s = \$102$. We want to test the hypotheses: $H_0 : \mu = 670$, $H_1 : \mu > 670$, using a significance level of $\alpha = 0.01$. Note that $n \geq 30$, which is a threshold above which we can consider $s \simeq \sigma$, and $\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim \mathcal{N}(0,1)$, approximately. Proceed as follows: **(a)** compute the "$z$-score", i.e. the value of the test statistic $z = (\overline{X} - \mu)/(s/\sqrt{n})$, under the null hypothesis $H_0$. **(b)** Find the acceptance and rejection regions $S_0$ and $S_1$ respectively. **(c)** Determine whether $z \in S_1$: should we reject the null hypothesis? **(d)** Finally, compute the $p$-value of the test statistic.

men earnings

$n = 40$

$\bar{x}$

$H_0: \mu = 670$

$H_1: \mu > 670$

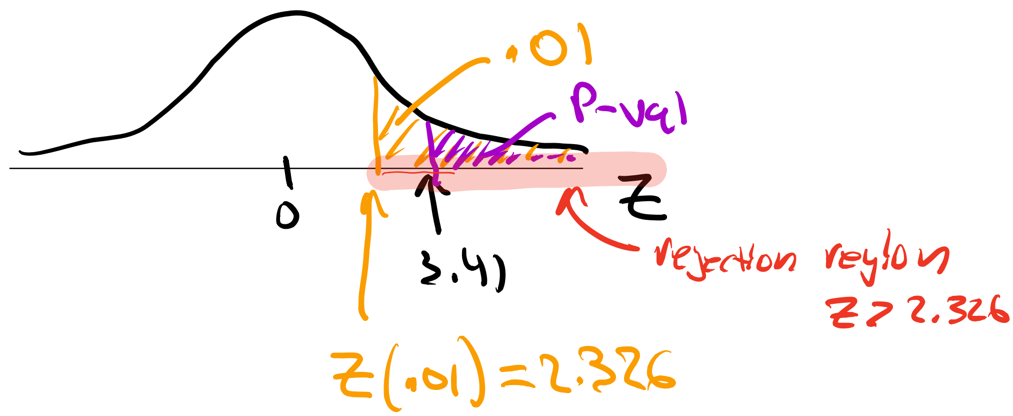
$\alpha = .01$

$\bar{x} = 725$

$S = 102$

$n = 40$

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{725 - 670}{102/\sqrt{40}} = 3.41$$




$3.41 > 2.316 \Rightarrow$ reject null

$\text{P-val} = 1 - \Phi(3.41) = .0003 \leq .01$

reject null.