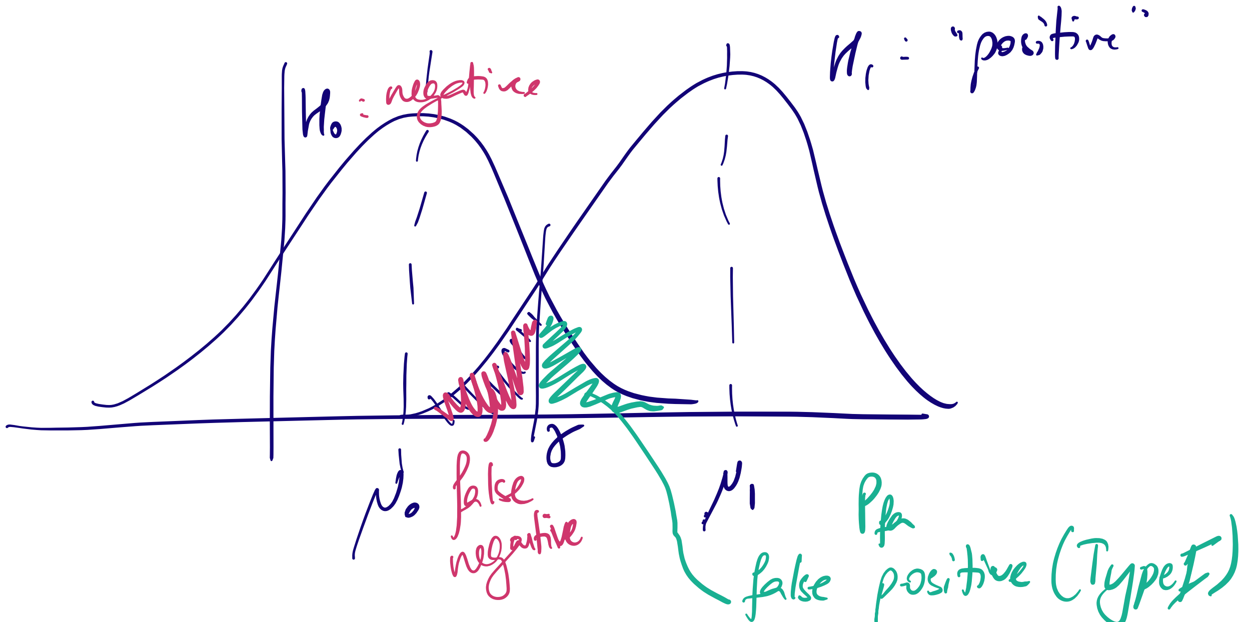
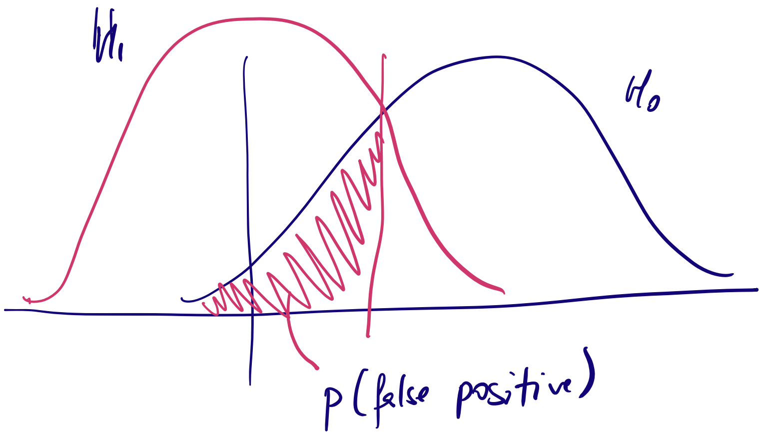
Lecture 19

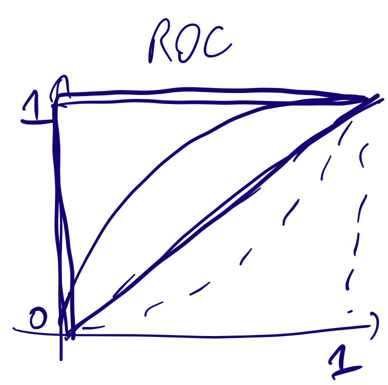
① Exam Review

② Linear Algebra!

①



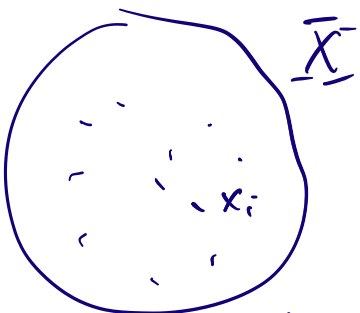




$0.5 \leq \text{AUC ROC} \leq 1$

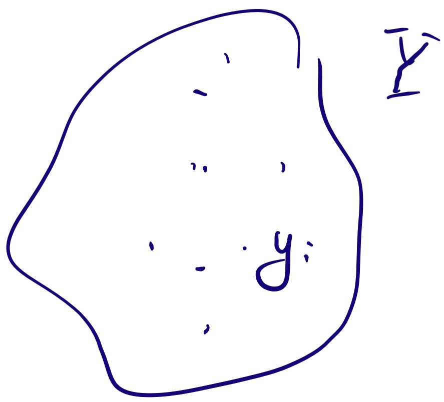
Hypothesis Testing

T-test / Z-test



firearm mortality from rural

vs.



firearm mortality from urban

sample mean for population $\bar{X}$

$$\hat{\mu}_x = \frac{1}{N_x} \sum_{i=1}^{N_x} x_i$$

sample mean from population $\bar{Y}$

$$\hat{\mu}_y = \frac{1}{N_y} \sum_{i=1}^{N_y} y_i$$

statistic

$$t = \hat{\mu}_x - \hat{\mu}_y$$

$\hookrightarrow$ instantiation of R.V. T

If $t = \hat{\mu}_x - \hat{\mu}_y \neq 0$?

$\hookrightarrow$ $\hat{\mu}_x \neq \tilde{\mu}_y$?

<u>Hypothesis Test:</u>

1) $H_0 : \mu_x = \mu_y$ $\qquad$ Two sets of data samples (X, Y)

$H_1 : \mu_x \neq \mu_y$

2) $H_0 : \mu_x = \mu_{provided}$ $\qquad$ One set of data samples

$H_1 : \mu_x \neq \mu_{provided}$

$\sigma_x^2 \equiv$ true variance for population X

$\sigma_y^2 \equiv$ " " " " Y

1) $T = \mu_X - \mu_Y = 0$ (under H_0)

Z-test

Known σ_X^2 & σ_Y^2

1) $T \sim G\left(0, \frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}\right)$

If $\sigma_X^2 = \sigma_Y^2 = \sigma^2$,

$T \sim G\left(0, \sigma^2\left(\frac{1}{N_X} + \frac{1}{N_Y}\right)\right)$

2) $T \sim G\left(0, \frac{\sigma^2}{N}\right)$

$T = \mu_X - \mu_{provided} = 0$

(under H_0)

T-test

Unknown σ_X^2 & σ_Y^2

$\Rightarrow$ Estimate the variance from data, e.g.

1) $S_X^2 = \frac{1}{N_X - 1} \sum_{i=1}^{N_X} (x_i - \bar{x})^2$

$S_Y^2 = \frac{1}{N_Y - 1} \sum_{i=1}^{N_Y} (y_i - \bar{y})^2$

$T \sim$ student's T dist.

(dof $= N_X + N_Y - 2$)

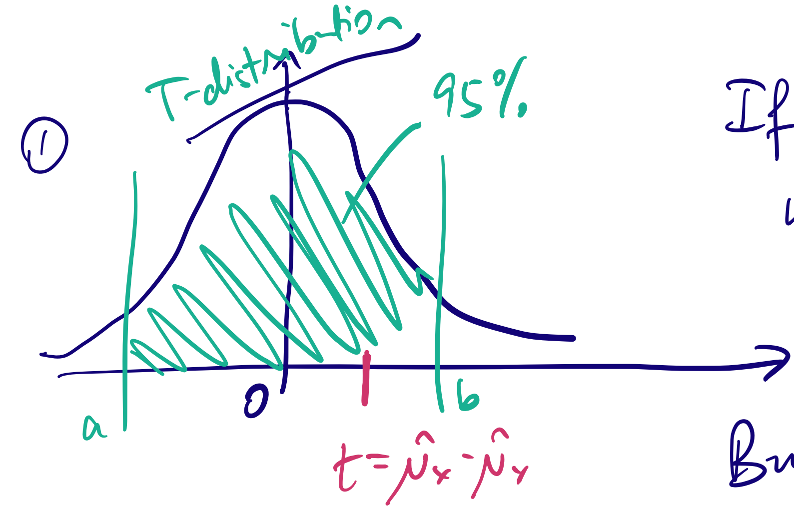
scale $= \sqrt{\sigma_X^2/N_X + \sigma_Y^2/N_Y}$

2) $T \sim$ T dist. (dof $= N-1$)

$t = \hat{\mu}_X - \hat{\mu}_Y$

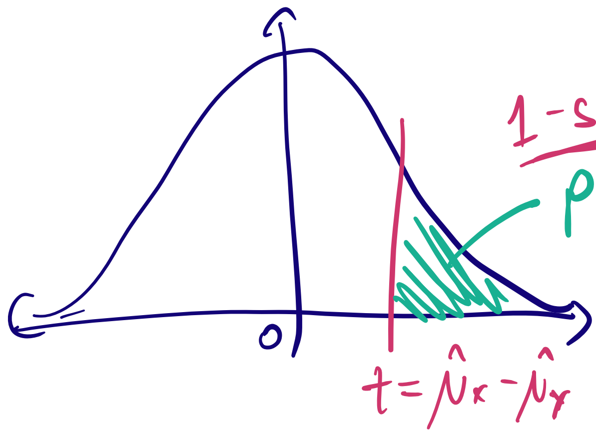
t is statistically significant :

①



If $t \in [a, b]$,
we cannot reject H_0

But if $t \notin [a, b]$,
we can reject H_0.



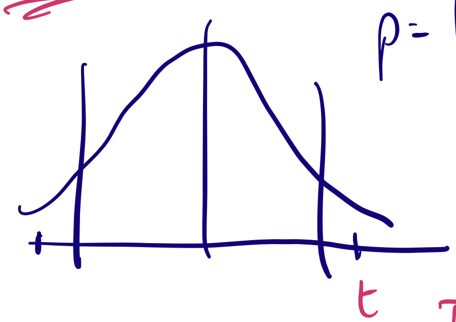
1-sided:

$p = \underline{P(T \geq t)}$ $\equiv$ survival function @ t

If $p < \alpha$ ($\alpha = 0.05$), then we reject H_0.

Otherwise, we cannot reject H_0.

2-sided:



$p = P(|T| \geq t)$, assuming $t \geq 0$

$p = P(|T| \geq t) = \underline{2\,P(T \geq t)}$

If $p < \alpha$, then we reject H_0.

If you're not sure, use a 2-sided hypothesis test.

1-sided:

H_1 : $\hat{\mu}_X > \hat{\mu}_Y$

$\hat{\mu}_X > \mu_{provided}$

CDF:

$$P(X \leq x) = \begin{cases} \sum_{k \leq x} p_X(k), & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{x} f_X(x)\,dx, & \text{if } X \text{ is continuous} \end{cases}$$

Goodness-of-fit measures

Discrete R.V.

χ^2-test

$T \sim \chi^2(K)$

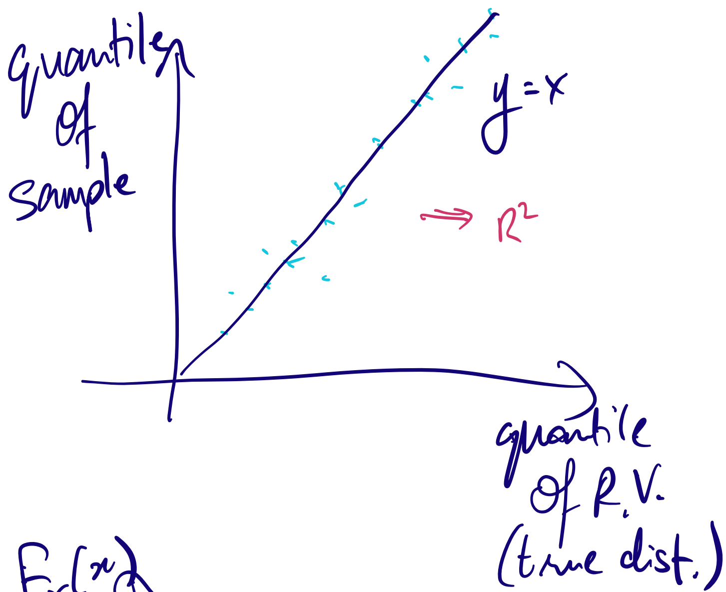
$\text{dof} = N_x - 1$

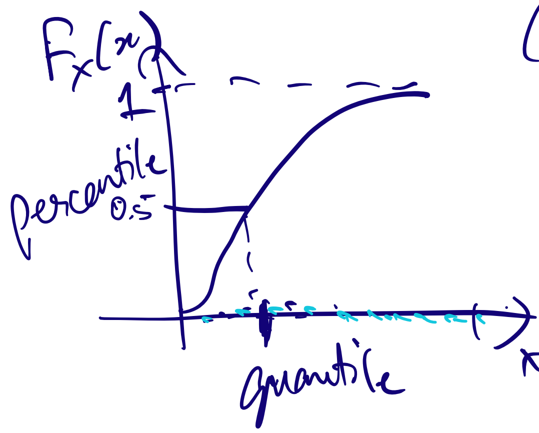
$$chi = \sum_{i=1}^{N_x} \frac{(x_i - E[x])^2}{E[x]}$$

$E[x] = \hat{\nu}_x$

$P(T \geq chi) = p$

Continuous

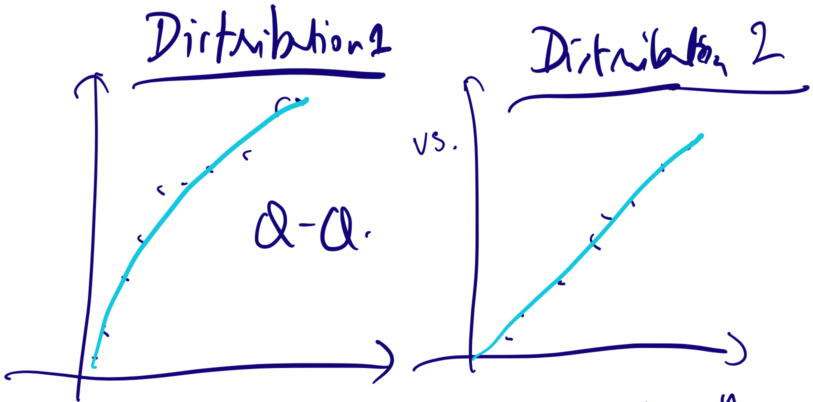




Q-Q plot

R^2 = coefficient of determination

$R^2 > 0.9$, it is a good fit



If $R_1^2 < R_2^2 \Rightarrow$ pick dist^n 2.