p-value

Under the assumption of the null hypothesis being true, a p-value is the probability of observing a result at least as extreme as the one observed by pure chance

a test stat

or even a statistic calculated from the sample (eg sample mean)

underlying premise

we believe in the real world that likely events occur.

End of the course

↓ review

Review

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample of n data points. Find the least squares estimates of

$$\hat{y}_i = \hat{\beta} x_i^2 \qquad i = 1, \ldots, n$$

relationship is quadratic, but terms are linear

Could do this directly (similar to OLS)

Slightly easier way is using a substitution

Let $y_i = x_i^2$. Estimator is

$$\hat{y}_i = \beta y_i, \quad i = 1, \ldots, n$$

Criteria for minimizing under least squares framework

Objective: minimize sum of squared residuals

obs y − fitted y

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \hat{y}_i = \hat{\beta} x_i$$

$$\min \sum (y_i - \hat{\beta} x_i^2)^2$$

Let $S(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i^2)^2$

$$\min \sum (y_i - \hat{\beta} x_i^2)^2$$

$$\min \underbrace{\sum (y_i - \hat{\beta} u_i)^2}$$

Let $u_i = x_i^2$

$$\frac{\partial S(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \sum (y_i - \beta u_i)^2$$

$$= \sum_{i=1}^{n} 2 (y_i - \beta u_i)(-u_i)$$

$$= \sum_{i=1}^{n} -2 (y_i u_i - \beta u_i^2)$$

Set $\partial S(\beta)/\partial \beta$ to zero, solve for β

$$\frac{\partial S(\beta)}{\partial \beta} = 0$$

$$\sum_{i=1}^{n} -2 (y_i u_i - \beta u_i^2) = 0$$

$$\sum_{i=1}^{n} y_i u_i - \beta \sum_{i=1}^{n} u_i^2 = 0$$

$$\beta = \frac{\sum_{i=1}^{n} y_i y_i^*}{\sum_{i=1}^{n} y_i^{*2}}$$

back substitute $y_i^* = x_i^2$ to return to terms in x

$$\beta = \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}$$

A researcher is testing whether a new teaching method changes test scores. The historical mean for the test has been 70 with a known standard deviation of 10 (continue to assume same St. dev for new method).

(a) State the appropriate null and alternative hypotheses of the researcher's test

$H_0: \mu = 70$ $H_a: \mu_a \neq 70$

(b) Using a sample of $n=25$ and a significance level of $\alpha = 0.05$, calculate the power of the test if the true mean is 74.

$n=25$, $\alpha = 0.05$, $\sigma = 10$, in reality $\mu = 74$

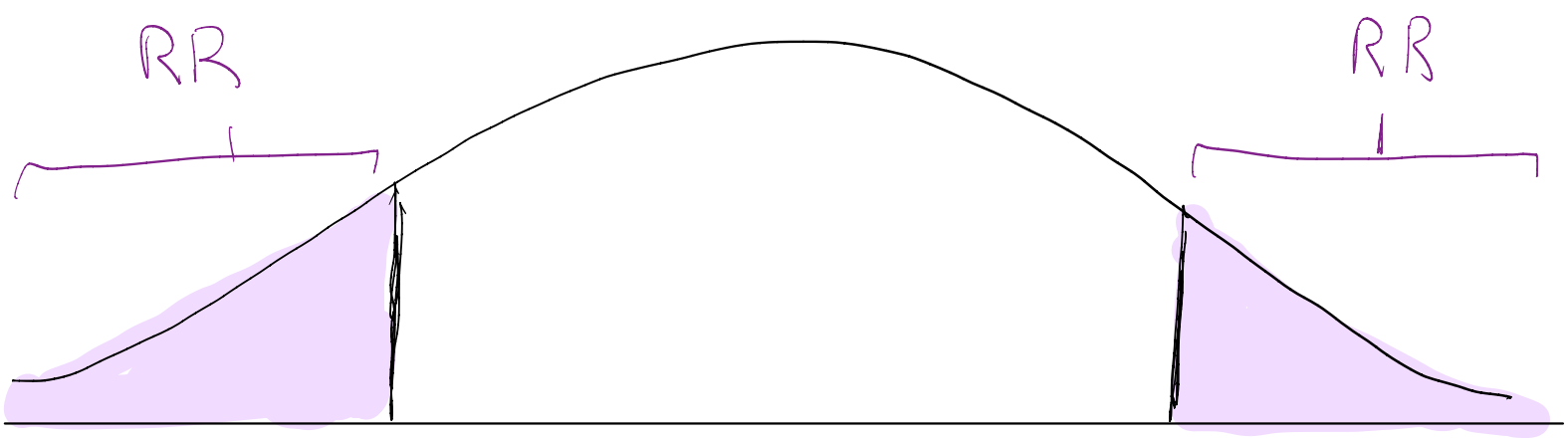
$H_0: \mu = 70$ $H_a: \mu \neq 70$ at $\alpha = 0.05$

Determine rejection region (RR)

$\sigma = 10$ known → use standard normal distrib

$\alpha = 0.05$ $H_a: \neq$

0.025 0.025



$-z_{crit}$ $+z_{crit}$

-1.96 $+1.96$

(normal table) (normal table)

From normal table: $z_{crit} = 1.96$

$-z_{crit} = -1.96$

Rejection Region

Reject H_0 if

(1) (2)

$z^* < -1.96$ or $z^* > 1.96$

test stat

using RR's, find equivalent $\bar{x}_{crit}$ under $H_0: \mu = 70$

RR ①

$$z^* < -1.96$$

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -1.96$$

$$\frac{\bar{x}_{crit} - 70}{10/\sqrt{25}} < -1.96$$

$$\bar{x}_{crit} < 66.08$$

RR ②

$$z^* > +1.96$$

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > +1.96$$

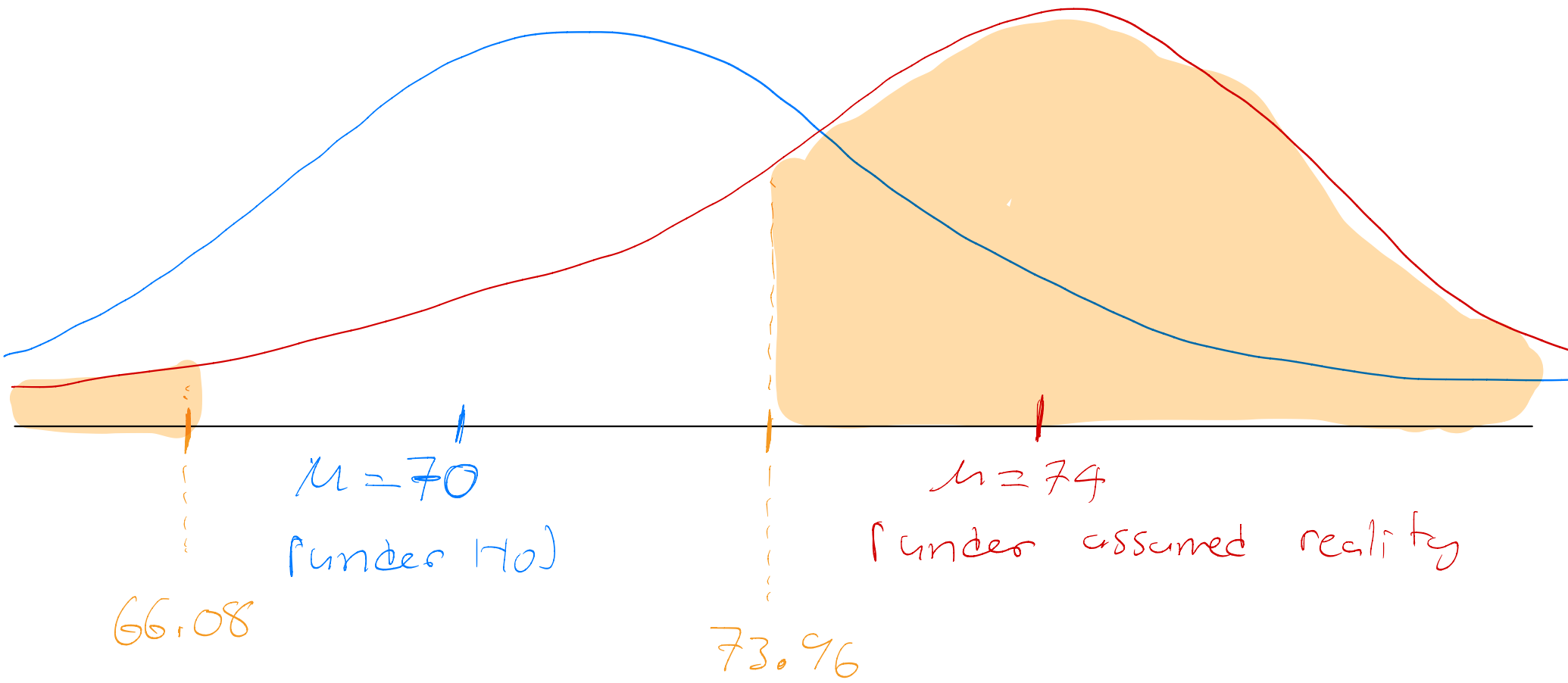
$$\frac{\bar{x}_{crit} - 70}{10/\sqrt{25}} > +1.96$$

$$\bar{x}_{crit} > 73.92$$

(Rejecting H_0 if $z^* < -1.96$ or $z^* > 1.96$)

$\Leftrightarrow$ (Rejecting H_0 if $\bar{x} < 66.08$ or $\bar{x} > 73.92$)

Calculate power under reality $\mu = 74$



power $= P(\bar{x} < 66.08 \mid \mu = 74) + P(\bar{x} > 73.92 \mid \mu = 74)$

$$= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{66.08 - \mu}{\sigma/\sqrt{n}} \,\middle|\, \mu = 74 \right)$$

($\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is Z)

$$+ P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{73.92 - \mu}{\sigma/\sqrt{n}} \,\middle|\, \mu = 74 \right)$$

($\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ is Z)

$$= P\left(Z < \frac{66.08 - 74}{10/\sqrt{25}} \right) + P\left(Z > \frac{73.92 - 74}{10/\sqrt{25}} \right)$$

$$= P(Z < -3.96) + P(Z > -0.04)$$

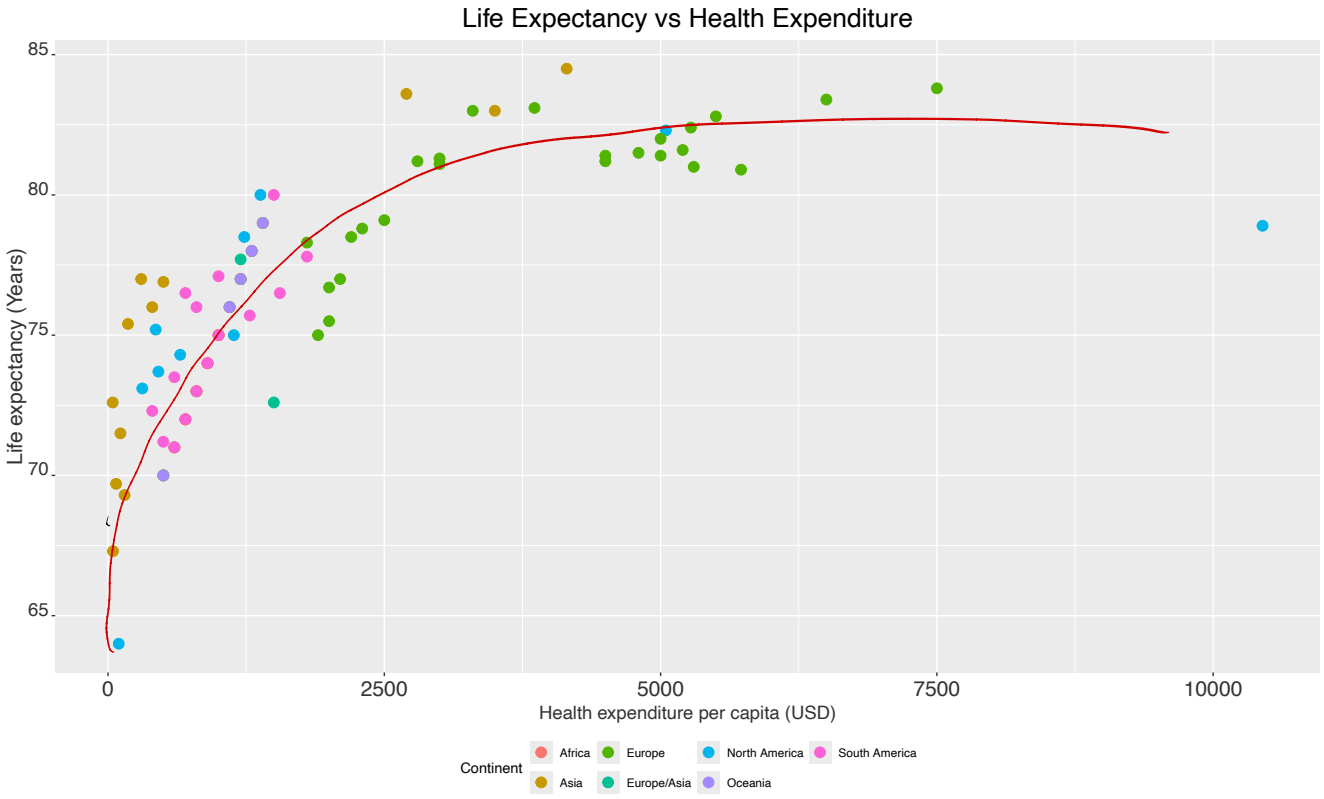
$P(Z < -3.96) \approx 0.00003$

$P(Z > -0.04) = 0.5160$

$$\approx 0.51603$$

power of this test to detect $\mu = 74$ is 51.603%

A plot of life expectancy vs health expenditure per capita is plotted below:



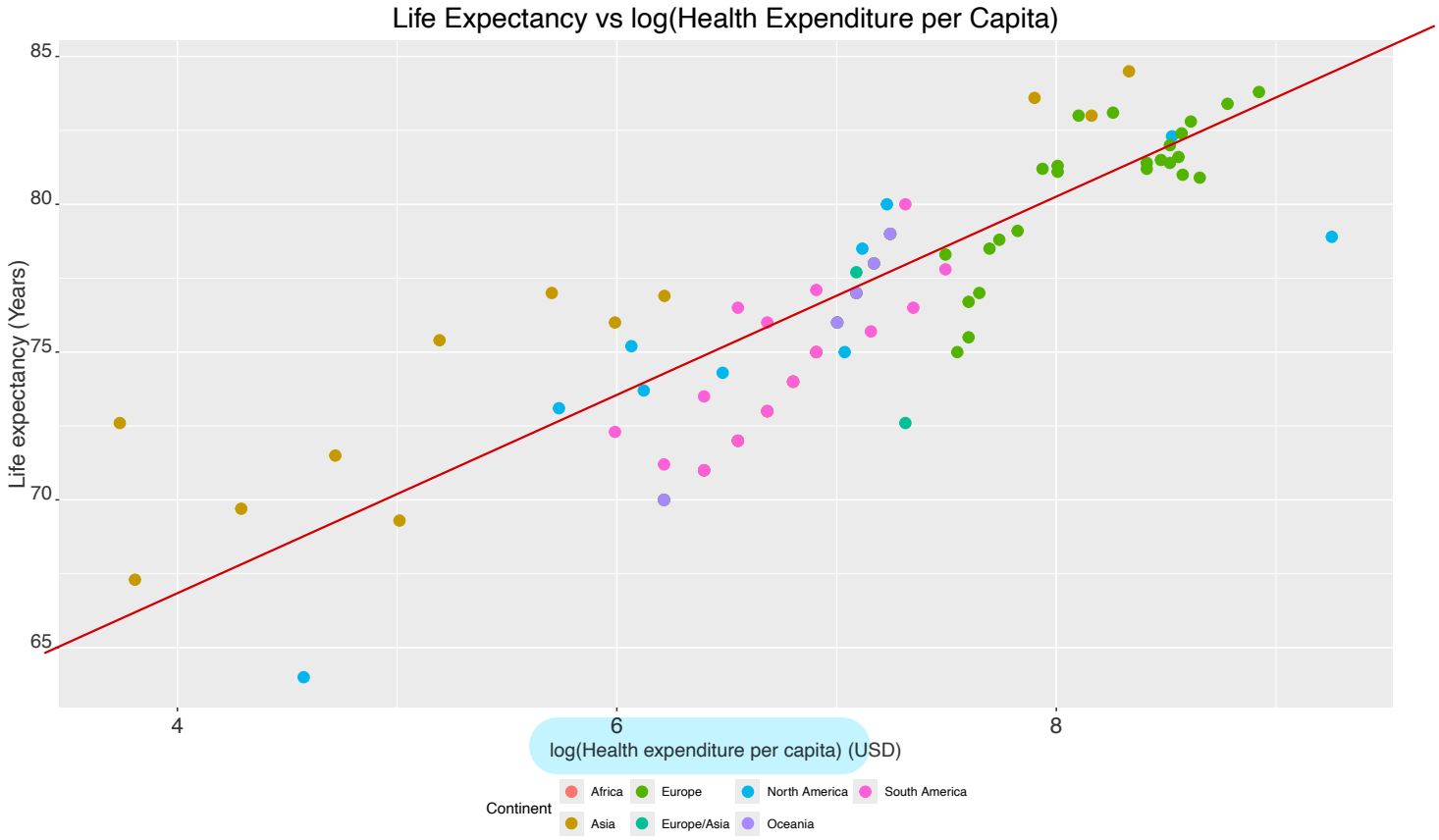
(a) Is a simple linear regression model appropriate for the data?

No.

A model with log transformed data is more appropriate.

$\log = \log_e = \ln$

(b) A plot of life expectancy vs. log(health expenditure) is given below



For the transformed data, the model was

$\hat{y} = 53.2740 + 3.2834x$

Interpret the slope

The model suggest an increase in log(health expenditure) by 1 logarithmic unit would result in an increase in life expectancy of 3.2834 years on average

(c) What is the life expectancy of a person born in a country which spends $5000 USD per capita on health expenditure

$x = \log(5000) = 8.517$

$\hat{y} = 53.2740 + 3.2834 x$

$= 53.2740 + 3.2834 (8.517)$

$= 81.239$ yrs.

A study was conducted to analyze daily protein intake of 6 groups of athletes

Group
1 Runners
2 Weight-lifters
3 Swimmers
4 Soccer players
5 Cyclists
6 Gymnists

$k = 6$ groups

(d) Complete the partially filled ANOVA table

SS / df

Source	df	Sum of Squares	Mean Square	F-Stat
Treatment/group	$k-1 = 5$	$SSTrt = 13580$	$MSTrt = 2716$	$F = 9.05$
Error	$n-k = 54$	$SSE = 16230$	$MSE = 300$	X
Total	$n-1 = 59$	$SSTotal = 29810$	X	X

$SSTrt + SSE = SSTotal$

$SSTrt + 16230 = 29810$

$SSTrt = 13580$

k = # groups $= 6$

$MSTrt = \frac{13580}{5} = 2716$

$MSE = \frac{16230}{54} = 300$

$F = \frac{MSTrt}{MSE} = \frac{2716}{300} = 9.05$

common variance

$s_p^2 = \hat{\sigma}^2 = MSE = 300$

(b) Use the ANOVA table to conduct the relevant hypothesis test. Use $\alpha = 0.01$.

$H_0: \mu_1 = \mu_2 = \dots = \mu_6$

H_a: At least one μ_i $i = 1, \dots, 6$ is different

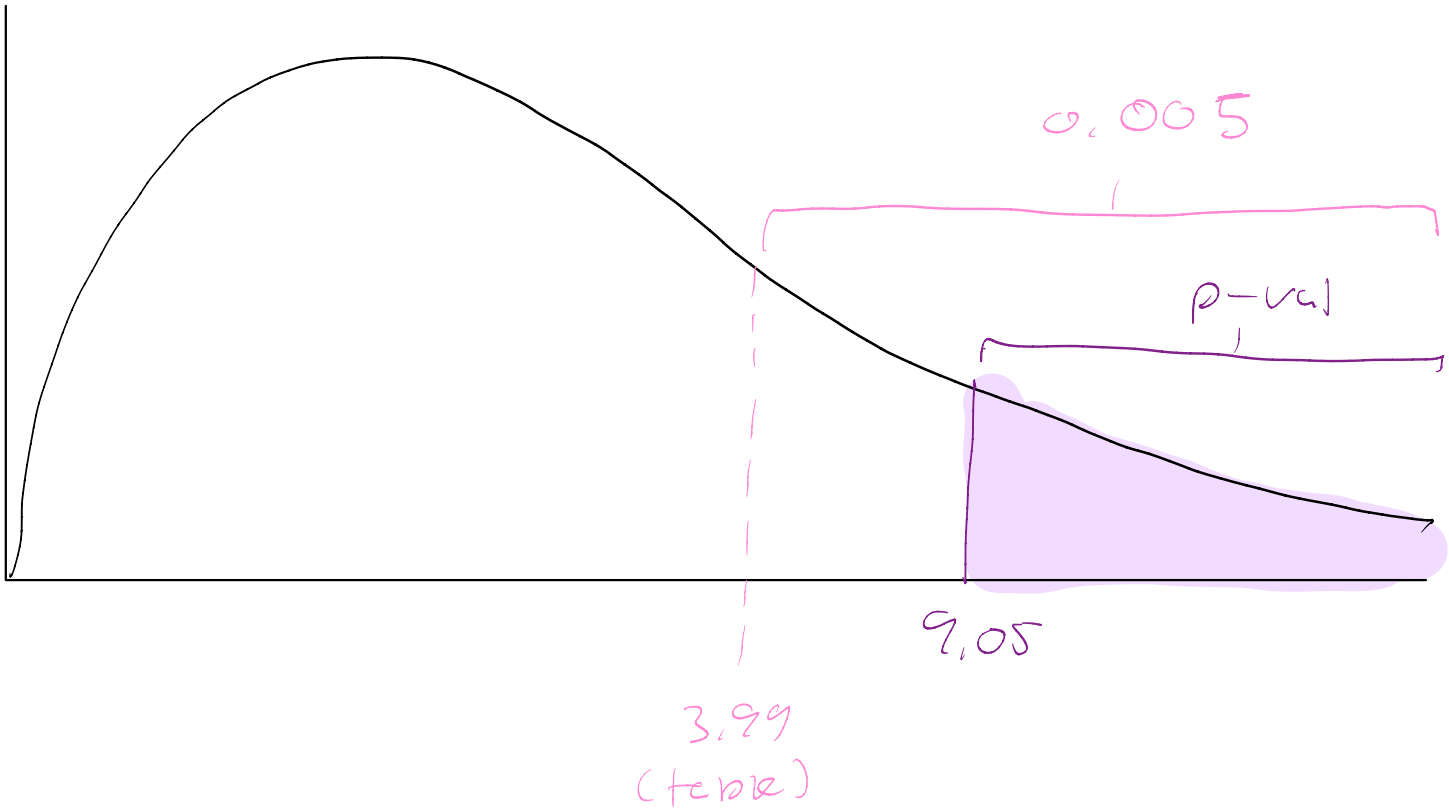
test stat: $F^* = 9.05$

Reference dist

F distribution with numerator = 5
and denominator = 54

use 40 df in F-table

F at 5 and 40 df



p-val < 0.005 < 0.01
(a)

Sufficient evidence to reject the null hypothesis and conclude the mean protein intake of at least one group of athletes is different

(c) Summary table

Group		Sample mean	Sample St.dev	n
(runners)	1	$\bar{X}_1 =$ 92	10	$n_1 =$ 12
	2	130	12	10
	3	110	9	11
	4	120	14	9
(cyclists)	5	$\bar{X}_5$ 98	11	$n_5 =$ 10
	6	105	13	8

Calculate a pairwise 95% CI using Fisher's LSD for the difference in means between cyclists and runners

⑤ ①

$1-\alpha = 0.95$

$\alpha = 0.05$

$$(\bar{X}_5 - \bar{X}_1) \pm t_{(n-k,\ \alpha/2)} \sqrt{\underbrace{S_p^2}_{MSE} \left(\frac{1}{n_5} + \frac{1}{n_1}\right)}$$

$$= (98 - 92) \pm t_{(54,\ 0.025)} \sqrt{300 \left(\frac{1}{10} + \frac{1}{12}\right)}$$

$t_{(54,\ 0.025)}$ ↓ use ∞ df (1.96)

$$= 6 \pm 14.536$$

$$= (\ -\ ,\ +\ )$$

Interp:

0 inside, plausible means equal

Instead of LSD, what if we did Bonferroni?

pairwise comparisons $\quad m = \binom{k}{2} = \binom{6}{2} = 15$

$$(\bar{x}_5 - \bar{x}_1) \pm t_{(n-k,\ \alpha/2m)} \sqrt{\underbrace{s_p^2}_{MSE} \frac{1}{n_5} + \frac{1}{n_1}}$$

$n - k = 54$

$\alpha/2m \downarrow$

$$\frac{0.05}{(2)(15)}$$

$t_{(54,\ 0.0016)}$

use ∞

not in t-table?

use closest rounded up.

0.005

$t(\infty, 0.005)$

$= 2.576$