Lecture 11

Topics: Hypothesis testing.

When estimating an unknown θ from data there are generally two types of statements you can make:

1) θ is likely within $[a, b]$ (comes from the data) } Confidence intervals

2) θ is likely not bigger than b or smaller than a } Hypothesis testing.

The two are effectively equivalent but often 2) can be done more convincingly.

e.g. Paper in PNAS 2012, Moss-Racusin et al.

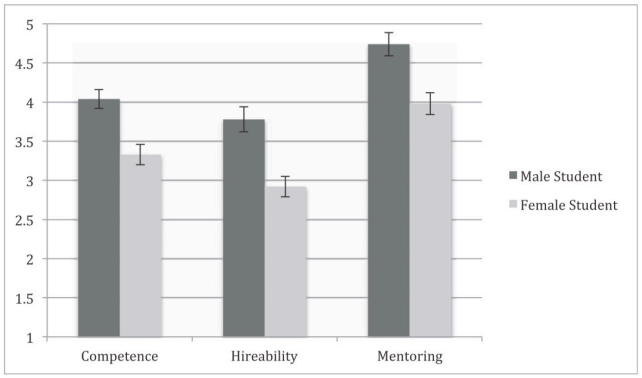


Fig. 1. Competence, hireability, and mentoring by student gender condition (collapsed across faculty gender). All student gender differences are significant ($P < 0.001$). Scales range from 1 to 7, with higher numbers reflecting a greater extent of each variable. Error bars represent SEs. $n_{\text{male student condition}} = 63$, $n_{\text{female student condition}} = 64$.

Sample of 127 faculty split into two groups:

group 1

63 faculty given a grad student application with a Male name and were asked to rate the student (scale 1-7)

group 2

64 faculty given the same student application but the name was changed to a female name. Asked to rate the student (scale 1-7)

group 1

Competence ratings: $X_1, X_2, \dots, X_{63}$

$\bar{X} = 4.1$

s.d. $= 1.19$

group 2

Competence ratings: $Y_1, Y_2, \dots, Y_{64}$

$\bar{Y} = 3.3$

s.d. $= 1.1$

Let μ_M = ave competence rating for Male name if shown to all faculty

μ_F = ave competence rating for Female name if shown to all faculty

The parameter of interest is $\mu_M - \mu_F$, in particular if $\mu_M - \mu_F = 0$ ← called the null hypothesis

In the data $\bar{X} - \bar{Y} > 0$ but, could this be due to sampling variability alone?

Lets start with a statement of type 1).

Note first that bias$(\bar{X}-\bar{Y}) = 0$ since

$$E(\bar{X}-\bar{Y}) = E(\bar{X}) - E(\bar{Y})$$

$$\overset{MF1}{=} \mu_M - \mu_F$$

← if the sample of faculty was drawn with replacement from all faculty

Now $$sd(\bar{X}-\bar{Y}) \overset{MF2}{=} \sqrt{var(\bar{X}) + var(\bar{Y})}$$

$$= \sqrt{\frac{var(X_1)}{63} + \frac{var(Y_1)}{64}}$$

$$\approx \sqrt{\frac{1.19^2}{63} + \frac{1.1^2}{64}}$$

← from the s.d.s in the two groups.

$$= 0.2$$

∴ Before collecting the data have

≈ 68% chance: $(\bar{X}-\bar{Y})$ within $(\mu_M - \mu_F) \pm 0.2$

≈ 95% chance: $(\bar{X}-\bar{Y})$ within $(\mu_M - \mu_F) \pm 2(0.2)$

≈ 99% chance: $(\bar{X}-\bar{Y})$ within $(\mu_M - \mu_F) \pm 3(0.2)$

Therefore

≈ 68% chance: $(\mu_M - \mu_F)$ within $(\bar{X}-\bar{Y}) \pm 0.2$

≈ 95% chance: $(\mu_M - \mu_F)$ within $(\bar{X}-\bar{Y}) \pm 2(0.2)$

≈ 99% chance: $(\mu_M - \mu_F)$ within $(\bar{X}-\bar{Y}) \pm 3(0.2)$

After collecting the data we have

$$\bar{X}-\bar{Y} = 4.1 - 3.3 = 0.8$$

Plugging this into the right hand side of the intervals above gives approximate confidence intervals.

≈ 68% CI: $(\mu_M - \mu_F)$ within $0.8 \pm 0.2 = (0.6, 1.0)$

≈ 95% CI: $(\mu_M - \mu_F)$ within $0.8 \pm 2(0.2) = (0.4, 1.2)$

≈ 99% CI: $(\mu_M - \mu_F)$ within $0.8 \pm 3(0.2) = (0.19, 1.4)$

type 1) statement

One problem with these approximate CI's is that they do not account for the fact that 0.1899 came from a random draw.

Next lecture we will see how to fix this.

A somewhat more effective argument can be made by a statement of type 2)

Temporarily suppose

$\mu_M - \mu_F = 0$ or even $\mu_M - \mu_F \leq 0$

called the null hypothesis, H_0

Now the z-score of $\bar{X}-\bar{Y}$ is

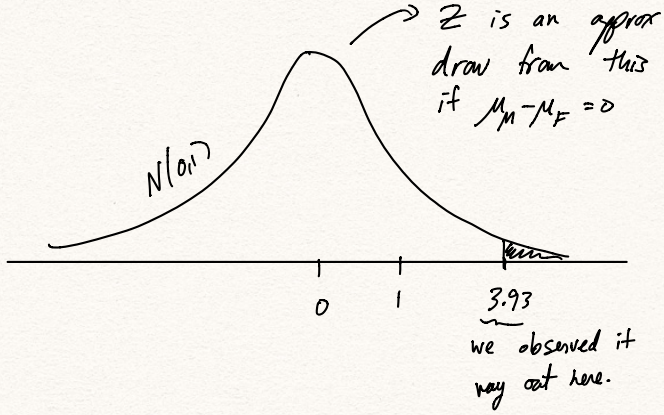
$$Z = \frac{(\bar{X}-\bar{Y}) - E(\bar{X}-\bar{Y})}{sd(\bar{X}-\bar{Y})} = \frac{(\bar{X}-\bar{Y}) - (\mu_M-\mu_F)}{\sqrt{\frac{var(X)}{63} + \frac{var(Y)}{64}}}$$

If we are assuming $\mu_M - \mu_F = 0$ then

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_M-\mu_F)}{\sqrt{\frac{var(X)}{63} + \frac{var(Y)}{64}}}$$

from the data → $= \frac{0.8 - 0}{\sqrt{\frac{var(X)}{63} + \frac{var(Y)}{64}}}$ ← null hypothesis

$$\approx \frac{0.8 - 0}{\sqrt{\frac{1.19^2}{63} + \frac{1.1^2}{64}}} = 3.93$$



Note $P(Z > 3.93) = 4 \times 10^{-5}$ ... this is unlikely and probably due to an incorrect assumption $\mu_M - \mu_F = 0$.

$P(Z > 3.93) = 4 \times 10^{-5}$ is an approx p-value for testing the null hypothesis

If we are assuming $\mu_M - \mu_F \leq 0$ then

$$Z = \frac{(\bar{X}-\bar{Y}) - (\mu_M-\mu_F)}{\sqrt{\frac{var(X)}{63} + \frac{var(Y)}{64}}}$$

The least rare this could be is

$$\frac{0.8 - 0}{\sqrt{\frac{var(X)}{63} + \frac{var(Y)}{64}}} \approx \frac{0.8 - 0}{\sqrt{\frac{1.19^2}{63} + \frac{1.1^2}{64}}} = 3.93$$

The only problem with the conclusion

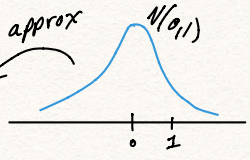
"$\bar{X}-\bar{Y}=0.8$ is inconsistent with $\mu_M=\mu_F$ since $P(Z>3.93)=4\times10^{-5}$ is too small to be explained by chance"

is that the z-score of 3.93 assumes $\hat{\sigma}_X=1.19$ and $\hat{\sigma}_Y=1.1$ are the true population values.

The traditional way to handle this is by adding some assumptions. To make the formulas more general write $n=63$ & $m=64$ (the number of samples in each group). Then

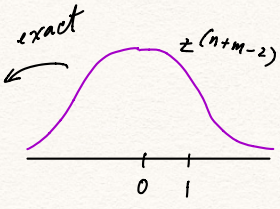
$$Z=\frac{(\bar{X}-\bar{Y})-(\mu_M-\mu_F)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}$$

$$\approx \frac{(\bar{X}-\bar{Y})-(\mu_M-\mu_F)}{\sqrt{\frac{\hat{\sigma}_X^2}{n}+\frac{\hat{\sigma}_Y^2}{m}}}$$



an approx draw from $N(0,1)$ if n & m are large and $\mu_M-\mu_F$ is the true value

$$\approx \frac{(\bar{X}-\bar{Y})-(\mu_M-\mu_F)}{\sqrt{\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^2}{m}}}$$



an exact draw from a $t^{(n+m-2)}$ distribution if:

- $X_1, X_2, \dots X_n \sim N(\mu_M, \sigma^2)$
- $Y_1, Y_2, \dots, Y_m \sim N(\mu_F, \sigma^2)$

} same variance

- $\hat{\sigma}^2=\frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{m}(Y_i-\bar{Y})^2\right)$

called the pooled estimate of sample variance.

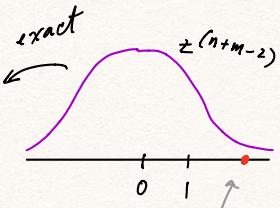
The pooled estimate from this data can be found as follows:

$$\frac{1}{63-1}\sum_{i=1}^{63}(X_i-\bar{X})^2=1.19^2,\quad n=63$$

$$\frac{1}{64-1}\sum_{i=1}^{64}(Y_i-\bar{Y})^2=1.1^2,\quad m=64$$

$$\therefore \hat{\sigma}^2=\underbrace{\frac{1}{n+m-2}}_{=\frac{1}{62+63}}\left[\underbrace{\sum_{i=1}^{n}(X_i-\bar{X})^2}_{=62(1.19)^2}+\underbrace{\sum_{i=1}^{m}(Y_i-\bar{Y})^2}_{=63(1.1)^2}\right]=1.31$$

$$\therefore Z\approx\frac{(\bar{X}-\bar{Y})-(\mu_M-\mu_F)}{\sqrt{\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^2}{m}}}$$



$$=\frac{0.8}{\sqrt{\frac{1.31}{63}+\frac{1.31}{64}}}\quad \text{if } \mu_M-\mu_F=0$$

$$=3.938$$

Now if T is a r.v. with a $t^{(63+64-2)}$ dist

Then $P(T>3.938)=\underbrace{0.0001}_{\text{half as rare as } 4.5\times10^{-5}}$

So we can either ...

i) approximate $Z = \dfrac{\bar{X}-\bar{Y}-(\mu_M-\mu_F)}{\sqrt{\frac{\hat{\sigma}_x^2}{n}+\frac{\hat{\sigma}_y^2}{m}}} \sim N(0,1)$

ii) assume $X_i \sim N(\mu_M, \sigma^2)$, $Y_i \sim N(\mu_F, \sigma^2)$

So that $T = \dfrac{\bar{X}-\bar{Y}-(\mu_M-\mu_F)}{\sqrt{\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^2}{m}}} \sim t^{(n+m-2)}$

... and conclude

"The difference in average rating between group 1 and group 2 is statistically significant (p-value < 0.001)"

... and is simply shorthand for:

"If we suppose $\mu_M = \mu_F$ then the z-score of $\bar{X}-\bar{Y} = 0.8$ is excessively large. Indeed the change of getting a z-score as large, or larger, is less than 0.001. Since it is hard to concive this was due to pure chance, we conclude it is likely that $\mu_M \neq \mu_F$ and, in particular, $\mu_M > \mu_F$"

Note!!! We have not discussed the magnitude of $\mu_M - \mu_F$ at all, just that its likely that $\mu_M > \mu_F$. However the magnitude is extremely important for making conclusions & is often overlooked

Another example:

- Warren Buffets daily expenditures over the last 10 years : 72.15$, 50.10$, ...
- my daily expenditures over the last 10 years : 26.25$, 42.15$, ...

Randomly sample daily recipts:

Warren Buffet	Ethan
$X_1, X_2, \ldots, X_{25}$	$Y_1, Y_2, \ldots, Y_{49}$
$\bar{X} = 177.92$ \$	$\bar{Y} = 151.11$ \$
$\hat{\sigma}_x^2 = 15.10^2$	$\hat{\sigma}_y^2 = 25.23^2$

Does this data suggest Warren has had larger average daily expenditure over the past 10 years?

null hypothesis H_0: $\mu_W = \mu_E$ (warrens 10 year ave = my 10 year ave)

alternative hypothesis H_A: $\mu_W > \mu_E$

To test this first note that

$$\hat{\sigma}^2 = \frac{1}{25+49-2}\left[\underbrace{\sum_{i=1}^{25}(X_i-\bar{X})^2}_{24\cdot(15.1)^2} + \underbrace{\sum_{i=1}^{49}(Y_i-\bar{X})^2}_{48(25.23)^2}\right] = 500.37$$

$\hat{\sigma}^2$: pooled est of variance

Therefore

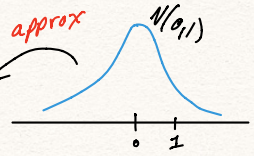
$$\sqrt{\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^2}{m}} = \sqrt{\frac{500.37}{25}+\frac{500.37}{49}} = 5.49$$

$$\sqrt{\frac{\hat{\sigma}_x^2}{n}+\frac{\hat{\sigma}_y^2}{m}} = \sqrt{\frac{15.1^2}{25}+\frac{25.23^2}{49}} = 4.7$$

If H_0 is true then $\bar{X} - \bar{Y} = 177.92 - 151.11 = 26.8$ ¢ has an approx Z score of

$$Z \approx \frac{(\bar{X}-\bar{Y}) - 0}{\sqrt{\frac{\hat{\sigma}_x^2}{n} + \frac{\hat{\sigma}_y^2}{m}}} \quad \swarrow \mu_W - \mu_E \text{ if } H_0$$

$$= \frac{26.8}{4.7} = 5.7$$



No way! Indeed

$$P\text{-value} = P(Z > 5.7) = 6 \times 10^{-9}$$

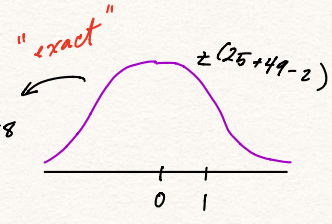
↑ $N(0,1)$; ↑ 6 in 1,000,000,000

So H_0 must be false & $\mu_W > \mu_E$.

If your willing to assume the two lists of daily recipts have a perfectly Normal histogram with equal variances, then

$$Z \approx \frac{(\bar{X}-\bar{Y}) - 0}{\sqrt{\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2}{m}}} \quad \swarrow \mu_W - \mu_E \text{ if } H_0$$

$$= \frac{26.8}{5.49} = 4.88$$



No way! Indeed

$$P\text{-value} = P(T > 4.88) = 3 \times 10^{-6}$$

change.

↑ $t^{(25+49-2)}$; ↑ 3 in 1,000,000

So H_0 must be false & $\mu_W > \mu_E$.

Remark: The assumptions required for the "t-test" are most likely false. However, since it gives a more conservative, i.e. less dramatic, result it is generally preferred.

Remark: One reason I do not like p-values is that they are highly sensative to small changes in Z-scores. i.e.

$P(Z > 3) = 0.0013$ ← 1 in 1000

$P(Z > 4) = 3 \times 10^{-5}$ ← 3 in 100,000

In some sense Z-scores are on the right scale for measuring rarity, the p-value scale is too dramatic and is sensative to assumptions or CLT approximations.

∴ I would prefer people just learn to interpret/calibrate rarity from Z-scores:

$|Z| \leq 3$ ⟺ Noting to get excited about

$3 < |Z| \leq 4$ ⟺ Something fishy is going on

$|Z| > 4$ ⟸ The estimate is inconsistent with the null hypothesis (or the assumptions of the experiment).

Regression

Regression is the workhourse of statistics.

A recent tweet:

"OLS: usually wrong, rarely bested"

ordinary least squares i.e. regression.

The idea is as follows:

- There is some variable Y and you want to see if another variable X can "explain" the variability in Y.
- Postulate a linear model

$$Y = \alpha + \beta X + Z$$

where Z is independent of X & $Z \sim N(0, \sigma^2)$.

- For a given set of X values

$$X_1, X_2, \dots, X_n$$

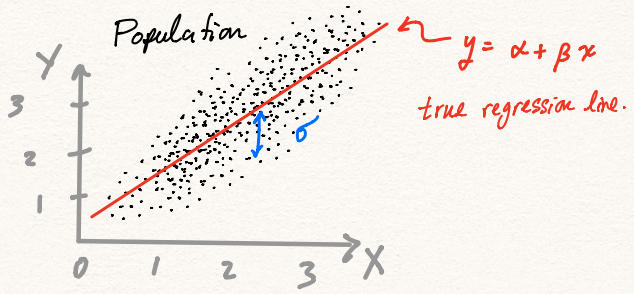
you measure the associated Y values

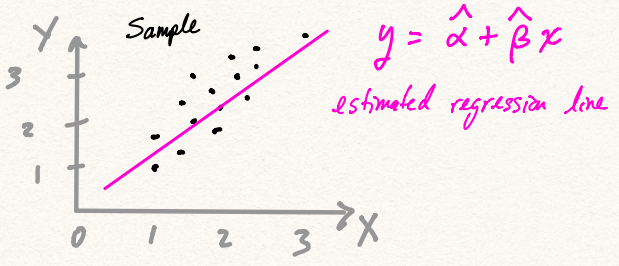
$$Y_1, Y_2, \dots, Y_n$$

- These two lists are used to construct an estimate $\hat{\beta}$ of β.
- Now $\hat{\beta}$ can be used to test

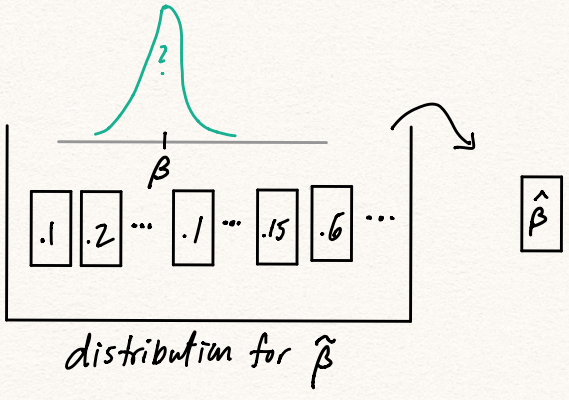
$H_0: \beta = 0$ ← which means Y has no linear relationship with X.

Here is the picture:





To do any inference (hypothesis tests or confidence intervals) we need to figure out the sampling variability of $\hat{\beta}$



Estimates of β, α, σ^2

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - (\hat{\alpha} + \hat{\beta} X_i)\right)^2$$

Sampling variability of $\hat{\beta}$ given $X_1, X_2, \ldots, X_n$ are fixed (i.e. conditional on $X_1, \ldots, X_n$):

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right)$$

plug in $\hat{\sigma}$ for an estimate of $sd(\hat{\beta})$

The estimate of β looks complicated but it is actually natural once you know the relationship btwn β & correlation ρ.

Recall $\beta = \rho \frac{sd(Y)}{sd(X)}$

$$= \frac{cov(X,Y)}{sd(X)\, sd(Y)} \frac{sd(Y)}{sd(X)}$$

$$= \frac{cov(X,Y)}{var(X)}$$

Notice

$$cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

$$= E\left[(X - E(X))Y\right] + E\left[(X - E(X))(-E(Y))\right]$$

$$\overset{MF1}{=} E\left[(X - E(X))Y\right] - E(Y)\underbrace{E\left[(X - E(X))\right]}_{=0}$$

$$= E\left[(X - E(X))Y\right]$$

$$\therefore \beta = \frac{E\left[(X - E(X))Y\right]}{E\left[(X - E(X))^2\right]} \quad \begin{matrix} \leftarrow cov(X,Y) \\ \leftarrow var(X) \end{matrix}$$

$$\approx \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Now fixing $X_1, X_2, \ldots, X_n$ as non-random

$$var(\hat{\beta}) = var\left(\frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right)$$

$$\overset{MF2}{=} \frac{1}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^2} var\left(\sum_{i=1}^{n} (X_i - \bar{X}) Y_i\right)$$

$$\overset{MF2}{=} \frac{1}{\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right]^2} \left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \underbrace{var(Y_i)}_{=\sigma^2}\right)$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Example:

Interview 5 individuals from California.

Ask them "How many years in school" & "How much did you make last year?"

years in school X	annual earnings Y
$X_1 = 16$	$38{,}000 = Y_1$
$X_2 = 14$	$21{,}000 = Y_2$
$X_3 = 11$	$14{,}000 = Y_3$
$X_4 = 16$	$25{,}000 = Y_4$
$X_5 = 18$	$30{,}000 = Y_5$

For the model:

$$Y = \alpha + \beta X + z, \quad z \overset{\text{indp.}}{\sim} N(0, \sigma^2)$$

we get...

$$\hat{\beta} = \frac{76{,}000}{28} \simeq 2{,}714$$

$$\hat{\alpha} = 25{,}600 - \hat{\beta}(15) \simeq -15{,}114$$

$$\hat{\sigma}^2 = 6400^2$$

... from the data

Is this data evidence that $\beta > 0$?

Certainly $\hat{\beta} = 2{,}714 > 0$ but does this suggest $\beta > 0$? Note

$$E(\hat{\beta}) = \beta$$

$$sd(\hat{\beta}) = \sqrt{var(\hat{\beta})}$$

$$= \sqrt{\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}}$$

$$\simeq \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}} = 1209.$$

$\therefore$ Assuming $\beta = 0$ the z-score of our obs of $\hat{\beta}$ is

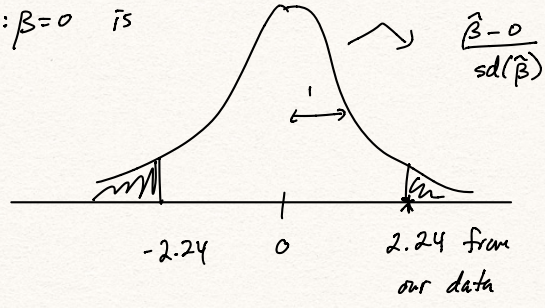
$$\frac{\hat{\beta} - E(\hat{\beta})}{sd(\hat{\beta})} = \frac{\hat{\beta} - 0}{sd(\hat{\beta})}$$

$$\simeq \frac{2714}{1209} = 2.24$$

pretty typical fluctuation of $N(0,1)$

$\therefore$ The data is consistent with $\beta = 0$ & doesn't conclusively say that β must be > 0

An approximate 2 sided p-value for testing $H_0: \beta = 0$ is



p-value ≈ 0.025

An approximate 95% CI for β is

$$\left(\hat{\beta} - 2\,sd(\hat{\beta}),\ \hat{\beta} + 2\,sd(\hat{\beta})\right)$$

$$= (296,\ 5132)$$

Example:

Suppose

$Y = \alpha + \beta X + Z$, Z & X are indep and $Z \sim N(0, \sigma^2)$

Based on random samples $\begin{pmatrix} X_1 \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} X_{25} \\ Y_{25} \end{pmatrix}$ one obtained the following statistics:

$$\bar{X} = 0.0335$$

$$\bar{Y} = -0.7713$$

$$\sum_{i=1}^{25} (X_i - \bar{X})^2 = 27.3$$

$$\sum_{i=1}^{25} (X_i - \bar{X}) Y_i = -269.2$$

Q1: Find $\hat{\beta}, \hat{\alpha}$:

$$\hat{\beta} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \frac{-269.2}{27.3} = -9.86$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = -0.7713 - (-9.86)\,0.0335$$

$$= -0.44$$

Q2: If $\hat{\sigma}^2 = 5.7711$ approximate $sd(\hat{\beta})$.

$$sd(\hat{\beta}) = \sqrt{\frac{\sigma^2}{\sum (X_i - \bar{X})^2}}$$

$$\approx \sqrt{\frac{5.7711}{27.3}} \quad (5.7711 \leftarrow \hat{\sigma}^2)$$

$$= 0.46$$

Q3: Find an approx 95% CI for β

For β: $-9.86 \pm 2(0.46)$

$$= (-10.78, -8.94)$$

Q4: Is there evidence that $\beta < 0$

$$\frac{\hat{\beta} - E(\hat{\beta})}{sd(\hat{\beta})} \approx \frac{-9.86 - 0}{0.46} \quad \text{if } \beta = 0$$

$$= -21.43$$

So if $\beta = 0$, $\hat{\beta}$ would be -21.43 $sd(\hat{\beta})$'s away from what we expect.

This is basically impossible.

If $\beta > 0$ it would be even more rare.

$\therefore$ the data conclusively says $\beta < 0$!

Q5: If I found another sample $\begin{pmatrix} X_{26} \\ Y_{26} \end{pmatrix}$ & told you $X_{26} = 0.05$

what is your best guess for Y_{26}.

Since $Y_{26} = \alpha + \beta X_{26} + Z$

$$E(Y_{26} \mid X_{26} = 0.05) = \alpha + \beta(0.05)$$

$$\approx \hat{\alpha} + \hat{\beta}(0.05)$$

$$= -0.933$$

Also notice that

$$sd(Y_{26} \mid X_{26} = 0.05) = \sigma$$
$$\approx \hat{\sigma}$$
$$= \sqrt{5.771}$$

$\therefore$ We would predict Y_{26} to be

$$-0.933 \pm \sqrt{5.771}$$

warning! this does not take into account the uncertainty in the estimates $\hat{\beta}$ & $\hat{\alpha}$ we used to predict -0.933