1. (5 pts) In a clinical test of the drug Viagra, 734 men were treated with Viagra and 725 men (the control group) were given a placebo. In the treatment group, 117 men experienced headaches and in the control group 29 men experienced headaches.

Is there enough evidence to support the claim, at the 0.01 significance level, that headaches occur at a **higher rate** among men who take Viagra than among men who do not?

Follow all the usual steps of the appropriate Hypothesis test to justify your answer.

(*) Write p_V = proportion of men taking Viagra who suffer headaches and p_N = proportion of men not taking Viagra who suffer headaches. The claim is $p_V > p_N$, so...

(1) **Hypotheses:** $H_0 : p_V = p_N$ vs. $H_a : p_V > p_N$

(2) **Significance level:** $\alpha = 0.01$.

(3a) **Test statistic:** Standard normal...

$$z = \frac{(\hat{p}_V - \hat{p}_N) - \overbrace{(p_V - p_N)}^{H_0: =0}}{\sqrt{\frac{\overline{p}\overline{q}}{n_V} + \frac{\overline{p}\overline{q}}{n_N}}} = \frac{(\hat{p}_V - \hat{p}_N)}{\sqrt{\frac{\overline{p}\overline{q}}{n_V} + \frac{\overline{p}\overline{q}}{n_N}}},$$

where $\overline{p}$ = pooled proportion ($\overline{q} = 1 - \overline{p}$),

$$\overline{p} = \frac{x_V + x_N}{n_V + n_N},$$

x_V = number of men in Viagra group who suffered head aches, x_N = the number of men in the control group who suffered headaches, n_V and n_N are the corresponding sample sizes and $\hat{p}_V$ and $\hat{p}_V$ are the corresponding sample proportions.

(3b) **Observed test statistic:** $x_V = 117$, $x_N = 29$, $n_V = 734$, $n_N = 725$, so

$$\hat{p}_V = \frac{117}{734},\ \hat{p}_N = \frac{29}{725} \text{ and } \overline{p} = \frac{117+29}{734+725} = \frac{146}{1459}$$

and

$$z^* = \frac{\frac{117}{734} - \frac{29}{725}}{\sqrt{\frac{\frac{146}{1459}\cdot\frac{1313}{1459}}{734} + \frac{\frac{146}{1459}\cdot\frac{1313}{1459}}{725}}} \approx 7.598$$

(4) **Decision criterion:** This is a right-tailed test (because of the form of H_a) and the P-value is

$$P = \text{Prob}(z > z^*) = \text{Prob}(z > 7.598) < 0.0001$$

so we reject H_0, since $P < \alpha = 0.01$.

(Alternatively, $7.598 > z_{0.01} \approx 2.33$, so we reject H_0.)

(5) **Conclusion:** There is enough evidence to support the claim (at the 0.01 significance level) that headaches occur at a higher rate among men who take Viagra than among men who do not.

2. (5 pts) A memory/sorting test was given to college students who use marijuana *lightly*, and the same test was given to an independent sample of college students who use marijuana *heavily*. The results are summarized in the table below.

Group	sample size	mean score	standard deviation
Light users	$n_1 = 64$	$\overline{x}_1 = 53.3$	$s_1 = 3.6$
Heavy users	$n_2 = 65$	$\overline{x}_2 = 50.3$	$s_2 = 4.5$

Test the claim at the $\alpha = 0.01$ significance level that the mean test score of light users of marijuana is higher than the mean test score of heavy users.

Follow all the usual steps of the appropriate Hypothesis test to justify your answer.

(*) Write μ_1 = mean score for light users of marijuana (in the population) and μ_2 = mean score for heavy users of marijuana. The claim is $\mu_1 > \mu_2$, so...

(1) **Hypotheses:** $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 > \mu_2$

(2) **Significance level:** $\alpha = 0.01$.

(3a) **Test statistic:** t-distribution with $m = \min(n_1, n_2) - 1 = 64 - 1$ d.f.

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \overbrace{(\mu_1 - \mu_2)}^{H_0:\,=0}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

(3b) **Observed test statistic:**

$$t^* = \frac{53.3 - 50.3}{\sqrt{\frac{3.6^2}{64} + \frac{4.5^2}{65}}} \approx 4.184$$

(4) **Decision:** This is a right-tailed test (because of the form of H_a) so we reject H_0 if $t^* > t_{0.01}$. In this case, $t_{0.01} \approx 2.39$ (using 60 d.f. to estimate for 64 d.f.), and we reject H_0 because $4.184 > 2.39$.

(5) **Conclusion:** There is enough evidence to support the claim (at the 0.01 significance level) that light users of marijuana score higher on average than heavy users.

3. A random sample of 54 bears was anesthetized and the widths of their heads and their weights were measured. These measurements and the relation between them are summarized in the table below.

Measurement	mean	standard deviation	correlation coefficient
Head width	$\overline{x} = 6.19$ inches	$s_x = 1.5$ inches	
Weight	$\overline{y} = 183$ pounds	$s_y = 4.5$ pounds	$r_{xy} = 0.78$

(a) (5 pts) Is there *significant* linear correlation between head width and weight? Justify your answer in terms of an appropriate hypothesis test at the $\alpha = 0.01$ significance level.

(*) Write $\rho =$ population correlation between head width and weight. So...

(1) $H_0 : \rho = 0$ vs. $H_a : \rho \neq 0$.

(2) $\alpha = 0.01$

(3a) $t = \dfrac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$, follows t-distribution with $n - 2 = 52$ d.f.

(3b) $t^* = \dfrac{0.78}{\sqrt{\dfrac{1-0.78^2}{52}}} \approx 8.988$.

(4) $8.988 > t_{0.01} \approx 2.403$ (using 50 d.f. to estimate for 52 d.f.), so we reject H_0.

(5) There is sufficient evidence to support the claim that their is significant linear correlation between weights of bears and the widths of their heads.

(b) (5 pts) The regression equation that explains the variation in bears' weights in terms of their head-widths is

$$\hat{y} = 168.52 + 2.34x.$$

(Based on the sample data and computed using the summary statistics above!)

What, according to this equation, is the predicted weight of a bear with a head-width of $x = 7.5$ inches? Is this prediction reliable? Why or why not?

The predicted weight is $\hat{y}(7.5) = 168.52 + 2.34 \cdot 7.5 \approx 186$ lbs. This estimate is reliable because (a) the correlation is significant so the linear relation is reliable, and (b) a head width of 7.5 inches is in the range of head widths in the data — less than 1 standard deviation above the sample average.

What, according to this equation, is the predicted weight of a bear with a head-width of $x = 1$ inch? Is this prediction reliable? Why or why not?

The predicted weight is $\hat{y}(1) = 168.52 + 2.34 \cdot 1 \approx 170$ lbs. This estimate is (much) less reliable because although there is significant correlation between the variables, a head width of 1 inch is well outside the likely range of head widths in the data — more than 3 standard deviation below the sample average.

Bonus (4 pts) Answer *true* or *false* to each of the claims below, and explain your answer briefly.

(i) *There is significant linear correlation between **years of education** and **income**. This means that going to school for longer will cause your income to rise.*

False: Association does ***not*** imply causation.

E.g., if a dentist goes back to school to study poetry for two years because she loves poetry, but does not use the additional schooling in her work, her income is not affected by the additional years of education.

(ii) *The linear correlation coefficient for paired data relating **monthly income** to **age** of people aged* 50 *to* 70 *is very close to* 0*. This means that, in this age group, there is no relation between age and income.*

False: If $r \approx 0$, then there is no ***linear*** relationship between age and income (in this age group), but there could well be some other relationship (e.g., quadratic).