Econometrics
Supervision 1

Samuel Lee

Question 1

(a)

The weights of the economists are normally distributed with a mean of 178 pounds and a standard deviation of 17 pounds. Denoting the weight of economist i as w_i, this means that $\frac{w_i-178}{17}$ follows a standard normal distribution with $\mu = 0, \sigma = 1$, and

$$P(w_i < x) = P\left(\frac{w_i - 178}{17} < \frac{x - 178}{17}\right) = \Phi\left(\frac{x - 178}{17}\right)$$

Thus the probability that any given economist weighs less than 189 pounds is $P(w_i < 189) = \Phi\left(\frac{189-178}{17}\right) = \Phi\left(\frac{11}{17}\right) \approx 0.7422$, and the probability that all the economists weigh less than 189 pounds is

$$\prod_{i=1}^{15} P(w_i < 189) \approx 0.7422^{15} = 0.01142$$

(b)

The probability that the raft is overloaded is

$$P\left(\sum_{i=1}^{15} w_i > 2850\right) = P\left(\frac{\sum_{i=1}^{15} w_i}{15} > \frac{2850}{15}\right) = P(\bar{w} > 190)$$

$\bar{w}$, the mean of the economists' weights, is normally distributed with mean of 178 pounds and a standard deviation of $\frac{17}{\sqrt{15}}$ pounds. Following the steps in (a),

$$\begin{aligned} P(\bar{w} > 190) = 1 - \Phi\left(\frac{190 - 178}{17/\sqrt{15}}\right) &= 1 - \Phi\left(\frac{12\sqrt{15}}{17}\right) \\ &\approx 1 - \Phi(2.734) \\ &\approx 1 - 0.99683 = 0.00317 \end{aligned}$$

(c)

For n economists that enter the raft, the probability of overloading is $P\left(\bar{w} > \frac{2850}{n}\right)$ and following (b), this is equal to

$$1 - \Phi\left(\frac{2850/n - 178}{17/\sqrt{n}}\right) = 1 - \Phi\left(\frac{2850 - 178n}{17\sqrt{n}}\right)$$

Since $1 - \Phi(x) < 0.0001$ when $x \gtrsim 3.731$, the probability of overloading is less than 0.0001 when

$$\begin{aligned}
\frac{2850 - 178n}{17\sqrt{n}} &\gtrsim 3.731 \\
2850 - 178n &\gtrsim 63.427\sqrt{n} \\
178n + 63.427\sqrt{n} - 2850 &\lesssim 0 \\
\sqrt{n} &\lesssim \frac{-63.427 + \sqrt{63.427^2 - 4(178)(-2850)}}{2 \cdot 178} \\
&\lesssim \frac{-63.427 + \sqrt{2033222.98433}}{356} \\
&\lesssim \frac{1425.911 - 63.427}{356} \\
&\lesssim 3.8272 \\
n &\lesssim 14.65
\end{aligned}$$

Thus the maximum number of economists that can enter to raft before the chance of overloading exceeds 0.0001 is 14.

Question 2

(a)

Figure 1: Histogram of wages

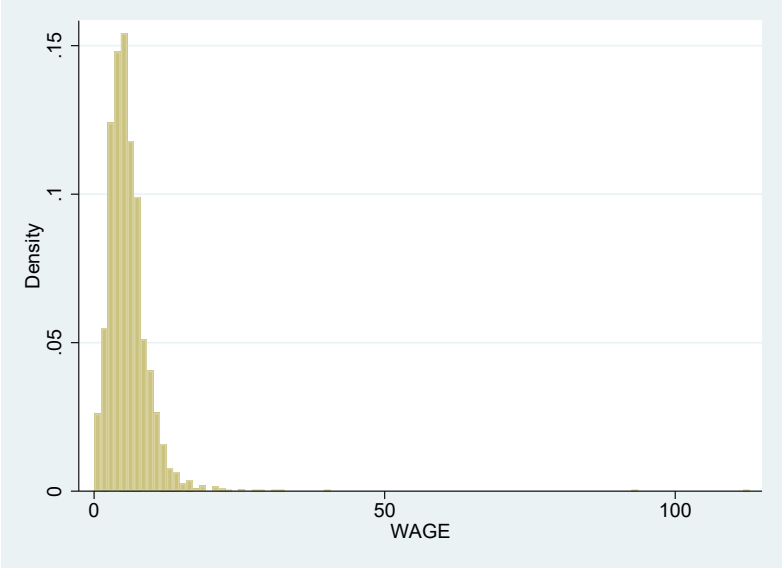
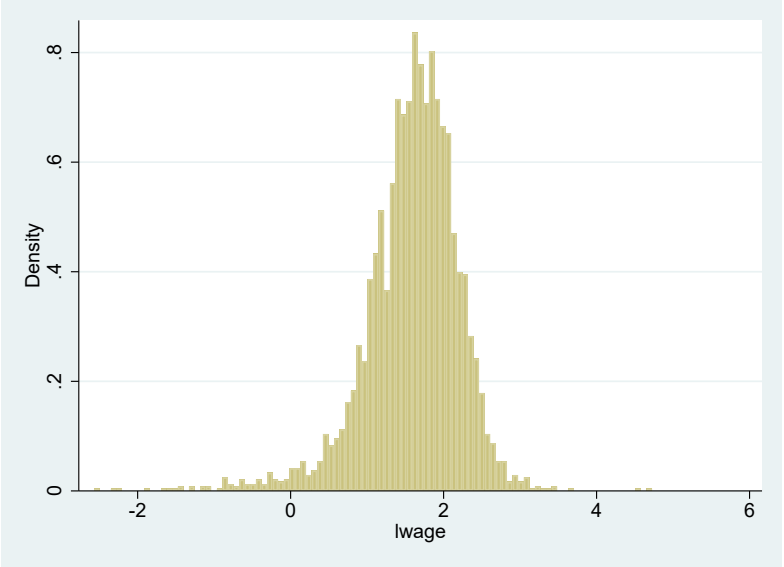


Figure 2: Histogram of logarithm of wages



The distribution for wages is skewed to the left, whereas the distribution for the logarithm of wages looks more symmetric and is closer to a normal distribution. When the logarithm of wages is taken, more weight is given to a unit increase when wages are low than when wages are higher $\left(\text{since } \frac{d^2 \ln x}{dx^2} = -\frac{1}{x^2} < 0\right)$. This 'compresses' the observations with very high wages and 'stretches' the observations that bunch around lower wages, normalizing the distribution.

(b)

Creating a variable for the logarithm of wages using the command *gen lwage = ln(wage)*, and using the command *sum lwage if male == 1*, the mean of the logarithm of wages is given as 1.693011 and the standard deviation is given as 0.6053204.

(c)

From (b), Stata calculates the estimate of the standard deviation of the logarithm of wages as $s = 0.6053204$. For some random variable with standard deviation σ, the estimated mean of this variable has a standard deviation of $\frac{\sigma}{\sqrt{n}}$ where n is the number of observations. This is because the standard deviation is the square root of the variance of the random variable. If the random variable has a variance of σ^2, the variance of all the observations added together is $Var(\sum_{i=1}^{n} w_i) = n\sigma^2$, and the variance of the mean is $Var\left(\frac{\sum_{i=1}^{n} w_i}{n}\right) = \frac{1}{n^2} Var(\sum_{i=1}^{n} w_i) = \frac{\sigma^2}{n}$. The standard deviation is then the square root which is $\frac{\sigma}{\sqrt{n}}$.

Using Stata's estimate of the standard deviation (which is the square root of the unbiased estimate of the population variance, and is itself not an unbiased estimate of the standard deviation due to Jensen's inequality), the mean of log-wage for males is $\frac{s}{\sqrt{3296}} = \frac{0.6053204}{\sqrt{3296}} = 0.01054367$.

(d)

The estimated mean of log-wage for males is normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, μ and σ being the population mean and standard deviation of the log-wage for males. Using the estimates Stata provides, the mean of the log-wage for males can be expected to follow a normal distribution with mean 1.693011 and standard deviation 0.01054367.

The assumption of normality for this distribution is justified by the central limit theorem, and is applicable when the number of observations is large (the condition is sometimes stated as $n > 40$).

To test that the mean of log wages for males is 1.7, the following hypotheses are considered:

$$H_0 : \mu = 1.7$$
$$H_1 : \mu \neq 1.7$$

The null hypothesis (H_0) is tested by first assuming that the null hypothesis is true, and then finding the ranges of values for which the probability of obtaining a sample statistic in those ranges were sufficiently low (below 5% by convention). If the sample statistic actually obtained in the current sample is in that range, one can reject the null hypothesis.

For a two-tailed test for inequality, if $P(\bar{w} < c_1 | H_0 \text{ is true})$ and $P(\bar{w} > c_2 | H_0 \text{ is true}) < 5\%$, and $\bar{w} < c_1$ or $\bar{w} > c_2$, the null hypothesis is rejected.

For this case, the critical values for the test statistic $\frac{w_i - 1.7}{0.01054367}$ are -1.96 and 1.96 (using the z-distribution since the number of observations is large enough for the t-distribution to approximate a Z-distribution). Since $\frac{w_i - 1.7}{0.01054367} = \frac{1.693011 - 1.7}{0.01054367} = -0.66286217$ which is between the critical values, the null hypothesis cannot be rejected.

(e)

The test is conducted as above, but the test statistic is now $\frac{u_m - u_f}{\sqrt{\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}}}$ since this is a two-sample test with the assumption of unequal variance. The test statistic is approximated as a t-distribution with the degrees of freedom being

$$\frac{\left(\frac{s_m^2}{n_m} + \frac{s_f^2}{n_f}\right)^2}{\frac{1}{n_m - 1}\left(\frac{s_m^2}{n_m}\right)^2 + \frac{1}{n_f - 1}\left(\frac{s_f^2}{n_f}\right)^2}$$

according to the Welch-Satterthwaite equation.

From Stata, $s_m^2 = 0.36641279$, $n_m = 1727$, $s_f^2 = 0.39787697$, and $n_f = 1569$. This means the degrees of freedom is calculated as ≈ 3233.2678 and a Z-distribution can be used as in (d).

With $\mu_m = 1.693011$ and $\mu_f = 1.474751$, the test statistic is

$$\frac{1.693011 - 1.474751}{\sqrt{\frac{0.36641279}{1727} + \frac{0.39787697}{1569}}} = 10.1133737717$$

Which is comfortably beyond any reasonable critical value one can take (taken to be 1.96 in the previous question).

(f)

No. For (d) one only needs to work with the distribution of the mean of log-wages for males, and thus the test is done assuming a normal distribution with the mean and standard deviation specific to males. To test that men and women have the same wage on average, one must take into account that the two groups have different distributions for their log-wages, and this will affect the probability that one uses to reject the null hypothesis. This is because μ_m and μ_f are themselves estimated by $\bar{w}_m$ and $\bar{w}_f$, and have their own variance about the true population means which will alter how stringent the test needs to be to reject H_0, as opposed to testing if μ_m is different from just some non-stochastic constant.

Question 3

(a)

This was calculated as part of 2(e): $s_f^2 = 0.39787697$.

(b)

The estimated variance has a χ^2_{n-1} distribution. More specifically, the statistic $\frac{(n-1)s^2}{\sigma^2}$ has a χ^2 distribution with n-1 degrees of freedom.

(c)

One could conduct an F-test, with the two hypotheses

$$H_0 : \frac{s_f^2}{s_m^2} = 1$$
$$H_1 : \frac{s_f^2}{s_m^2} \neq 1$$

where the test statistic is $\frac{s_f^2}{s_m^2}$ which follows an F-distribution with $n_f - 1$ and $n_m - 1$ degrees of freedom if H_0 is true.

The test statistic is $\frac{s_f^2}{s_m^2} = 1.0858709$, but the degrees of freedom are too large to be found on any statistical table (and taking (∞, ∞) makes the test meaningless). Using the Stata command *sdtest lwage, by(male)*, $Pr(F > f)$ is given as 0.0473 which means the null hypothesis could have been rejected at a 5% level for a one-tailed test, but should not be rejected for a two-tailed test.

(d)

There are many reasons why the variance for females might be expected to be larger. It could be that females undertake more part-time work due to family commitments, whereas full-time work is more common with males. There are also factors which will vary male wages more than female wages, for example if there is workplace discrimination that keeps female wages more bunched around the lower end. As it stands the data shows a higher variance for females than males.