Problem 1a.

$$\mathbb{E}[\mathbb{E}[Y|X,Z]|X] = \sum_{z=\{0,1\}} \mathbb{E}[Y|X=x,Z=z]\,P(Z=z|X=x)$$

$$= \sum_{z=\{0,1\}} \sum_{y} yP(Y=y|X=x,Z=z)\,P(Z=z|X=x)$$

$$= \sum_{z=\{0,1\}} \sum_{y} y\frac{P(Y=y,X=x,Z=z)}{P(X=x,Z=z)}\frac{P(Z=z,X=x)}{P(X=x)}$$

$$= \sum_{z=\{0,1\}} \sum_{y} y\frac{P(Y=y,X=x,Z=z)}{P(X=x)} = \sum_{y} yP(Y=y|X=x) = E[Y|X]$$

This is a more general version of the **law of iterated expectations**, which says that conditioning on more information first, $\{X, Z\}$, and then less information second, $\{Z\}$, is the same as only conditioning on the less information.

Problem 1b.

Using the law of iterated expectations and the fact that $\mathbb{E}[Y|X] = X^2$, we can show that

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big] = \mathbb{E}[X^2]$$

$$\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{E}[Y] - 0 = \mathbb{E}[Y]$$

Problem 1c.

Law of iterated expectations

$$\text{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] = \mathbb{E}\big[\mathbb{E}[XY|X]\big] = \mathbb{E}\big[X\mathbb{E}[Y|X]\big] = \mathbb{E}[X^3]$$

Therefore, when $\mathbb{E}[X^3] = 0$, X and Y are uncorrelated

One such example would be when $X \sim N(0,1)$ and $Y = X^2$

In that case, $\text{cov}[X, Y] = \mathbb{E}[X^3] = 0$, since this is the third moment ($\mu^3 + 3\mu\sigma^2$) of the standard normal distribution and it is 0. Therefore, X and Y are uncorrelated. However, since $\mathbb{E}[Y|X] = X^2$, X and Y will be dependent.

Another example would be when $X \sim U[-1,1]$ and $Y = X^2$

Problem 2a.

$$\mathbb{E}\,[(Y-c)^2] = \mathbb{E}[(Y-\mu+\mu-\ c)^2] = \mathbb{E}[(Y-\mu)^2 + 2(Y-\mu)(\mu-c) + (\mu-c)^2]$$

$$= \mathbb{E}[(Y-\mu)^2] + 2\,\mathbb{E}[(Y-\mu)(\mu-c)] + \mathbb{E}[(\mu-c)^2]$$

$$= \sigma_Y^2 + \underbrace{2(\mu^2 - c\mu - \mu^2 + c\mu)}_{=0} + (\mu-c)^2 = \sigma_Y^2 + (\mu-c)^2$$

When $c = \mu$, this gives the best prediction of Y in the sense of minimizing the mean-squared prediction error.

Problem 2b.

We need to show that $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \underset{\mathbf{h}(\mathbf{X})}{\mathbf{argmin}}\, \mathbb{E}\left[(\mathbf{Y} - \mathbf{h}(\mathbf{X}))^{\mathbf{2}}\right]$, where h(X) is some function of X.

$$(Y - h(X))^2 = (Y - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - h(X))^2$$

$$= (Y - \mathbb{E}[Y|X])^2 + \big(\mathbb{E}[Y|X] - h(X)\big)^2 + 2\big(\mathbb{E}[Y|X] - h(X)\big)(Y - \mathbb{E}[Y|X])$$

$(Y - \mathbb{E}[Y|X])^2$ does not matter for optimization since it does not include h(X).

$$\mathbb{E}\Big[\underbrace{\big(\mathbb{E}[Y|X] - h(X)\big)}_{b(X)}\underbrace{(Y - \mathbb{E}[Y|X]}_{\varepsilon}\Big] = \mathbb{E}[b(X)\varepsilon] = \mathbb{E}\big[\mathbb{E}[b(X)\varepsilon|X]\big] = \mathbb{E}\big[b(X)\mathbb{E}[\varepsilon|X]\big] = 0$$

Due to law of iterated expectations

Due to conditional mean independence

Therefore, $\big(\mathbb{E}[Y|X] - h(X)\big)^2$ will be minimized when $\mathbf{h}(\mathbf{X}) = \mathbb{E}[\mathbf{Y}|\mathbf{X}]$**.**

Problem 2c.

Definition of variance

Due to law of iterated expectations

$$\sigma_Y^2 = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \mathbb{E}\left[\mathbb{E}[Y^2|X]\right] - \left(\mathbb{E}\left[\mathbb{E}[Y|X]\right]\right)^2$$

Definition of conditional variance

$$\sigma_{Y|X}^2 = \mathbb{E}[Y^2|X] - (\mathbb{E}[Y|X])^2 \Rightarrow \mathbb{E}\left[\mathbb{E}[\mathbf{Y^2}|\mathbf{X}]\right] = \mathbb{E}\left[\sigma_{Y|X}^2\right] + \mathbb{E}[(\mathbb{E}[Y|X])^2]$$

Replacing $\mathbb{E}\left[\mathbb{E}[\mathbf{Y^2}|\mathbf{X}]\right]$ into the first equation will result in the following variance formula:

$$\sigma_Y^2 = \mathbb{E}\left[\sigma_{Y|X}^2\right] + \mathbb{E}[(\mathbb{E}[Y|X])^2] - \left(\mathbb{E}\left[\mathbb{E}[Y|X]\right]\right)^2 = \mathbb{E}\left[\sigma_{Y|X}^2\right] + \mathbb{V}\left[\mathbb{E}[Y|X]\right]$$

The law of total variance (also referred to as the analysis of variance or ANOVA) states that the variability of the random variable Y can be decomposed into two parts - the average variability "within" values of X and the variability "across" values of X. $\sigma_{Y|X}^2$ is "within-X" variance, i.e. variance in Y given X, while $\mathbb{V}\left[\mathbb{E}[Y|X]\right]$ is "between-X" variance, i.e. the variance of the CEF of Y given X.

Problem 2d.

$$\sigma_X^2 = \mathbb{E}[(X - \mu_X)^2]$$

$$= \mathbb{E}\left[(X - \mu_X)^2 \,\middle|\, |X - \mu_X| \geq c\sigma_X\right] P(|X - \mu_X| \geq c\sigma_X))$$

$$+\underbrace{\mathbb{E}\left[(X - \mu_X)^2 \,\middle|\, |X - \mu_X| < c\sigma_X\right]}_{\geq 0} \underbrace{P(|X - \mu_X| < c\sigma_X))}_{\geq 0}$$

$$= \mathbb{E}\left[(X - \mu_X)^2 \,\middle|\, |X - \mu_X| \geq c\sigma_X\right] P(|X - \mu_X| \geq c\sigma_X))$$

$$\geq c^2\sigma_X^2 \, P(|X - \mu_X| \geq c\sigma_X))$$

Dividing both sides by $c^2\sigma_X^2$ will result in Chebyshev's inequality: $\mathbf{P(|X - \mu_X| \geq c\sigma_X) \leq \frac{1}{c^2}}$

Problem 3a.

$$\mathbb{E}[\hat{\alpha}_1] = \mathbb{E}\left[\frac{n-1}{n}\bar{Y}_n\right] = \frac{n-1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\frac{Y_i}{n}\right] = \frac{n-1}{n^2}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{n-1}{n^2}n\mu = \frac{n-1}{n}\mu$$

$$\mathbb{E}[\hat{\alpha}_2] = \mathbb{E}\left[\frac{1}{2}\bar{Y}_n\right] = \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\frac{Y_i}{n}\right] = \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{2n}n\mu = \frac{1}{2}\mu$$

As $\mathbb{E}[\hat{\alpha}_1] \neq \mu$ and $\mathbb{E}[\hat{\alpha}_2] \neq \mu$, they are both biased estimators of μ , with the biases given as follows:

$$\text{Bias}[\hat{\alpha}_1, \mu] = \mathbb{E}[\hat{\alpha}_1] - \mu = \frac{n-1}{n}\mu - \mu = -\frac{\mu}{n}$$

$$\text{Bias}[\hat{\alpha}_2, \mu] = \mathbb{E}[\hat{\alpha}_2] - \mu = \frac{1}{2}\mu - \mu = -\frac{\mu}{2}$$

The bias of $\hat{\alpha}_1$ in absolute terms decreases when the sample size increases, while the bias of $\hat{\alpha}_2$ is independent of the sample size.

Problem 3b.

$$\mathrm{Var}[\hat{\alpha}_1] = \mathrm{Var}\left[\frac{n-1}{n}\overline{Y}_n\right] = \frac{(n-1)^2}{(n)^2}\mathrm{Var}\left[\sum_{i=1}^{n}\frac{Y_i}{n}\right] = \frac{(n-1)^2}{(n)^4}\sum_{i=1}^{n}\mathrm{Var}[Y_i] = \frac{(n-1)^2}{n^3}\sigma^2$$

$$\mathrm{Var}[\hat{\alpha}_2] = \mathrm{Var}\left[\frac{1}{2}\overline{Y}_n\right] = \frac{1}{4}\mathrm{Var}\left[\sum_{i=1}^{n}\frac{Y_i}{n}\right] = \frac{1}{4n^2}\sum_{i=1}^{n}\mathrm{Var}[Y_i] = \frac{1}{4n}\sigma^2$$

$$\mathrm{SE}[\hat{\alpha}_1] = \sqrt{\mathrm{Var}[\hat{\alpha}_1]} = \frac{(n-1)\sigma}{n\sqrt{n}}$$

$$\mathrm{SE}[\hat{\alpha}_2] = \sqrt{\mathrm{Var}[\hat{\alpha}_2] = \frac{\sigma}{2\sqrt{n}}}$$

$$\frac{\mathrm{SE}[\hat{\alpha}_1]}{\mathrm{SE}[\hat{\alpha}_2]} = \frac{(n-1)\sigma}{n\sqrt{n}} * \frac{2\sqrt{n}}{\sigma} = \frac{2(n-1)}{n}$$

When n>2, $\frac{2(n-1)}{n} > 1$ and $\hat{\alpha}_2$ is more precise than $\hat{\alpha}_1$.

On the other hand, when n>2, $|\mathrm{Bias}[\hat{\alpha}_1, \mu]| = \frac{\mu}{n} < \frac{\mu}{2} = |\mathrm{Bias}[\hat{\alpha}_2, \mu]|$, thus, $\hat{\alpha}_2$ is more biased.

Problem 3c.

$$\mathrm{MSE}[\hat{\alpha}_2] = \mathbb{E}[(\hat{\alpha}_2 - \mu)^2] = \mathrm{Bias}^2(\hat{\alpha}_2, \mu) + \mathrm{Var}(\hat{\alpha}_2) = \frac{\mu^2}{4} + \frac{\sigma^2}{4n}$$

$$\mathrm{MSE}[\bar{Y}_n] = \mathrm{Bias}^2(\bar{Y}_n, \mu) + \mathrm{Var}(\bar{Y}_n) = 0 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n}$$

$$\mathrm{MSE}[\hat{\alpha}_2] = \frac{\mu^2}{4} + \frac{\sigma^2}{4n} < \frac{\sigma^2}{n} = \mathrm{MSE}[\bar{Y}_n] \iff \boldsymbol{\mu^2 n < 3\sigma^2}$$

The basic intuition is that although $\bar{Y}_n$ is an unbiased estimator for μ, we might prefer the biased estimator $\hat{\alpha}_2$ because it gives a lower value of MSE compared to the unbiased estimator and requires a smaller sample size (less than $3\sigma^2/\mu^2$ in our example). This tradeoff is an extremely fundamental idea in machine learning, when we see that a small increase in bias can come with a large enough reduction in variance so that it decreases overall MSE. We might also prefer the biased estimator when we believe that the true population mean is zero, because even though on average we won't be correct, for any single instance of that estimator we'll be closer.

Problem 3d.

$$\mathbb{E}[\hat{\alpha}_k] = \mathbb{E}[k\bar{Y}_n] = k\mathbb{E}\left[\sum_{i=1}^{n} \frac{Y_i}{n}\right] = k\mu$$

$$\mathrm{Var}[\hat{\alpha}_k] = \mathrm{Var}[k\bar{Y}_n] = k^2\mathrm{Var}\left[\sum_{i=1}^{n} \frac{Y_i}{n}\right] = \frac{k^2\sigma^2}{n}$$

$$\mathrm{Bias}[\hat{\alpha}_k, \mu] = \mathbb{E}[\hat{\alpha}_k] - \mu = k\mu - \mu$$

$$\mathbf{MSE}[\hat{\boldsymbol{\alpha}}_{\mathbf{k}}] = \mathbf{Bias}^{\mathbf{2}}(\hat{\boldsymbol{\alpha}}_{\mathbf{k}}, \boldsymbol{\mu}) + \mathbf{Var}[\hat{\boldsymbol{\alpha}}_{\mathbf{k}}] = (\mathbf{k\mu} - \boldsymbol{\mu})^{\mathbf{2}} + \frac{\mathbf{k}^{\mathbf{2}}\boldsymbol{\sigma}^{\mathbf{2}}}{\mathbf{n}} \quad \rightarrow \mathbf{min}$$

The first order condition with respect to **k** will be:

$$2\mu(k\mu - \mu) + 2k\frac{\sigma^2}{n} = 0 \Rightarrow \mathbf{k}^* = \frac{\boldsymbol{\mu}^{\mathbf{2}}\mathbf{n}}{\boldsymbol{\mu}^{\mathbf{2}}\mathbf{n} + \boldsymbol{\sigma}^{\mathbf{2}}} = \frac{\boldsymbol{\mu}^{\mathbf{2}}}{\boldsymbol{\mu}^{\mathbf{2}} + \frac{\boldsymbol{\sigma}^{\mathbf{2}}}{\mathbf{n}}}$$

As the sample size increases, the sampling variance (σ^2/n) decreases and k^* approaches to 1, with the implication that the bias attenuates towards 0 and $k\bar{Y}_n$ becomes an unbiased estimator for μ. On the other hand, the larger is the population mean (μ), the larger is k^* that is required to minimize the MSE, since larger μ is associated with larger bias in absolute terms.

Problem 4a.

$$\mathbb{E}[\overline{Y}_n] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{Y_i}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] = \frac{1}{n} n[P(Y_i = 1) * 1 + P(Y_i = 0) * 0] = P(Y_i = 1) = \mu$$

Problem 4b.

$$S(Y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}[(Y_i)^2 - 2Y_i\overline{Y}_n + (\overline{Y}_n)^2]$$

$$= \frac{1}{n}\sum_{i=1}^{n}(Y_i)^2 - 2(\overline{Y}_n)^2 + (\overline{Y}_n)^2 = \overline{Y}_n - (\overline{Y}_n)^2 = \overline{Y}_n(1 - \overline{Y}_n)$$

$$\widehat{SE}(\overline{Y}_n) = \frac{S(Y_i)}{\sqrt{n}} = \sqrt{\frac{\overline{Y}_n(1-\overline{Y}_n)}{n}}$$

This formula is specifically for the case of a Bernoulli distribution, for which we can make use of the fact that $\sum_{i=1}^{n}(Y_i)^2 = \sum_{i=1}^{n} Y_i$. For other distributions, the formula of the sample standard deviation will be different.

Problem 4c.

According to Chebyshev's inequality that we proved earlier

$$P[|\overline{Y}_n - \mu| < \varepsilon] \geq 1 - \frac{Var(\overline{Y}_n)}{\varepsilon^2} = 1 - \frac{\sigma^2}{n\varepsilon^2}, \text{ where } \varepsilon \text{ is a positive number.}$$

Therefore, when $n \to \infty$, $\frac{\sigma^2}{n\varepsilon^2} \to 0$ and $P[|\overline{Y}_n - \mu| < \varepsilon] = 1$ or $\overline{Y}_n \xrightarrow{p} \mu$

Part B: Problem 1a.

If the experiment has no effect, i.e. the underlying population variances are equal between the treatment (**t**) and control (**c**) groups, the pooled standard error ($\mathbf{SE_{pooled}}$) of the difference in sample means can be calculated through the following formula:

$$\mathbf{SE_{pooled}} = \mathbf{S_{pooled}}\sqrt{\frac{1}{n_t}+\frac{1}{n_c}} = \sqrt{\frac{(n_t-1)S_t^2+(n_c-1)S_c^2}{n_t+n_c-2}}\sqrt{\frac{1}{n_t}+\frac{1}{n_c}}$$

where $\mathbf{S_i}$ is the sample standard deviation of group i, $\mathbf{n_i}$ – number of observations, $\mathbf{i} = \{\mathbf{t}, \mathbf{c}\}$**:**

On the other hand, if the experiment has an effect, i.e. the underlying population variances are not equal, the standard error of the difference in sample means can be derived as follows:

$$\mathbf{SE} = \sqrt{\frac{S_t^2}{n_t}+\frac{S_c^2}{n_c}}$$

Hypothesis testing for exhausted benefits: Claimant experiment (CE)

$$H_0:\ \mu_t^{CE} = \mu_c^{CE} \qquad H_1:\ \mu_t^{CE} \neq \mu_c^{CE}$$

Case 1: The experiment has no effect (pooled variance)

$$T_n = \frac{\overline{Y}_t^{CE} - \overline{Y}_c^{CE}}{SE_{pooled}^{CE}} =$$

$$= \frac{0.446 - 0.478}{\sqrt{\frac{(4186-1)*4186*\ 0.008^2 + (3952-1)*3952*0.008^2}{4186+3952-2}}\sqrt{\frac{1}{4186}+\frac{1}{3952}}} \approx -\frac{0.032}{0.011} \approx -2.90$$

If we assume a 5% significance level, the corresponding two-tailed critical value will be 1.96. As $|T_n| > |1.96|$, we reject the null hypothesis the Claimant Experiment had no effect on the proportion of UI claimants who exhausted their benefits.

Case 2: The experiment has an effect (different variances)

$$T_n = \frac{\overline{Y}_t^{CE} - \overline{Y}_c^{CE}}{SE^{CE}} = \frac{0.446 - 0.478}{\sqrt{0.008^2 + 0.008^2}} \approx -\frac{0.032}{0.011} \approx -2.90$$

As in the previous case, $|T_n| > |1.96|$, therefore, we reject the null hypothesis of no difference in means at the 5% significance level.

Hypothesis testing for exhausted benefits: Employer experiment (EE)

$$H_0: \ \mu_t^{EE} = \mu_c^{EE} \qquad H_1: \ \mu_t^{EE} \neq \mu_c^{EE}$$

Case 1: The experiment has no effect (pooled variance)

$$T_n = \frac{\overline{Y}_t^{EE} - \overline{Y}_c^{EE}}{SE_{pooled}^{EE}} =$$

$$= \frac{0.464 - 0.478}{\sqrt{\frac{(3963-1) * 3963 * \ 0.008^2 + (3952-1) * 3952 * 0.008^2}{3963 + 3952 - 2}} \sqrt{\frac{1}{3963} + \frac{1}{3952}}} \approx -\frac{0.014}{0.011} \approx -1.27$$

Again, if we assume a 5% significance level, the corresponding two-tailed critical value will be 1.96. However, since $|T_n| < |1.96|$, we fail to reject the null hypothesis that the Employer Experiment had no effect on the proportion of UI claimants who exhausted their benefits.

Case 2: The experiment has an effect (different variances)

$$T_n = \frac{\overline{Y}_t^{EE} - \overline{Y}_c^{EE}}{SE^{EE}} = \frac{0.464 - 0.478}{\sqrt{0.008^2 + 0.008^2}} \approx -\frac{0.014}{0.011} \approx -1.27$$

As in previous case, $|T_n| < |1.96|$, therefore, we fail to reject the null hypothesis that the Employer Experiment had no effect on the proportion of UI claimants who exhausted their benefits.

Part B: Problem 1b.

In general, the choice between a one-tailed or two-tailed test depends on the specific research question and hypothesis being tested. If the research question specifically focuses on whether the experiment reduced weeks of insured unemployment, then a one-tailed test might be more sensible as it is testing only for a decrease in weeks of insured unemployment and not an increase. However, if the research question is more general, such as whether the experiment had any effect on weeks of insured unemployment (either increase or decrease), then a two-tailed test would be more appropriate.

I think the two-tailed test is more appropriate in this case as we have no idea about direction of the effects on weeks of insured unemployment. For example, if UI benefits act as a subsidy to the consumption of nonmarket time (or leisure) and labor is not supplied perfectly inelastically, the availability of UI benefits may even increase unemployment duration instead of reducing it. Nevertheless, we will test both hypotheses in this section, assuming that the two samples come from different populations (in the previous section we have seen that this does not have a significant impact on the standard errors):

Hypothesis testing for weeks of insured unemployment: Claimant experiment (CE)

$$H_0: \ \mu_t^{CE} = \mu_c^{CE}$$

$$H_1: \ \mu_t^{CE} \neq \mu_c^{CE} \text{ or } H_1: \ \mu_t^{CE} < \mu_c^{CE}$$

$$T_n = \frac{\bar{Y}_t^{CE} - \bar{Y}_c^{CE}}{SE^{CE}} = \frac{17.0 - 18.3}{\sqrt{0.199^2 + 0.205^2}} \approx -\frac{1.3}{0.286} \approx -4.55$$

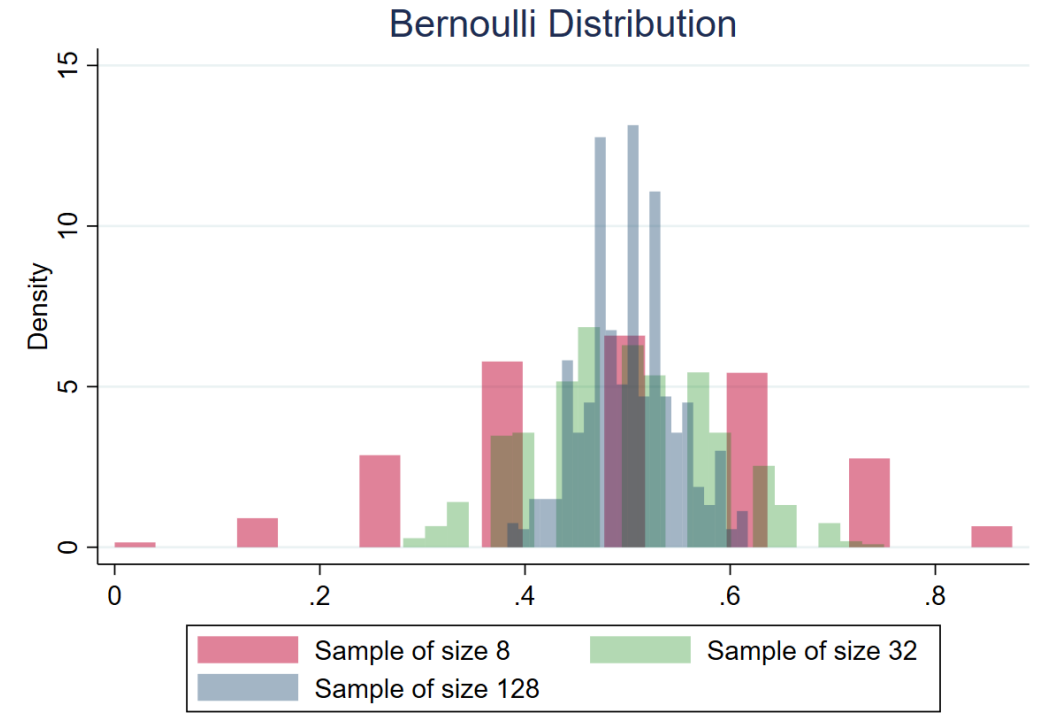
If we assume a 5% significance level, the corresponding one-tailed and two-tailed critical values will be -1.645 and 1.96 respectively. In the one-tailed test, we reject the null hypothesis that the experiment had no effect on the weeks of insured unemployment as $T_n = -4.55 < -1.645$. In the two-tailed test, we also reject the null hypothesis at the 5% significance level since $|T_n| = 4.55 > 1.96$.

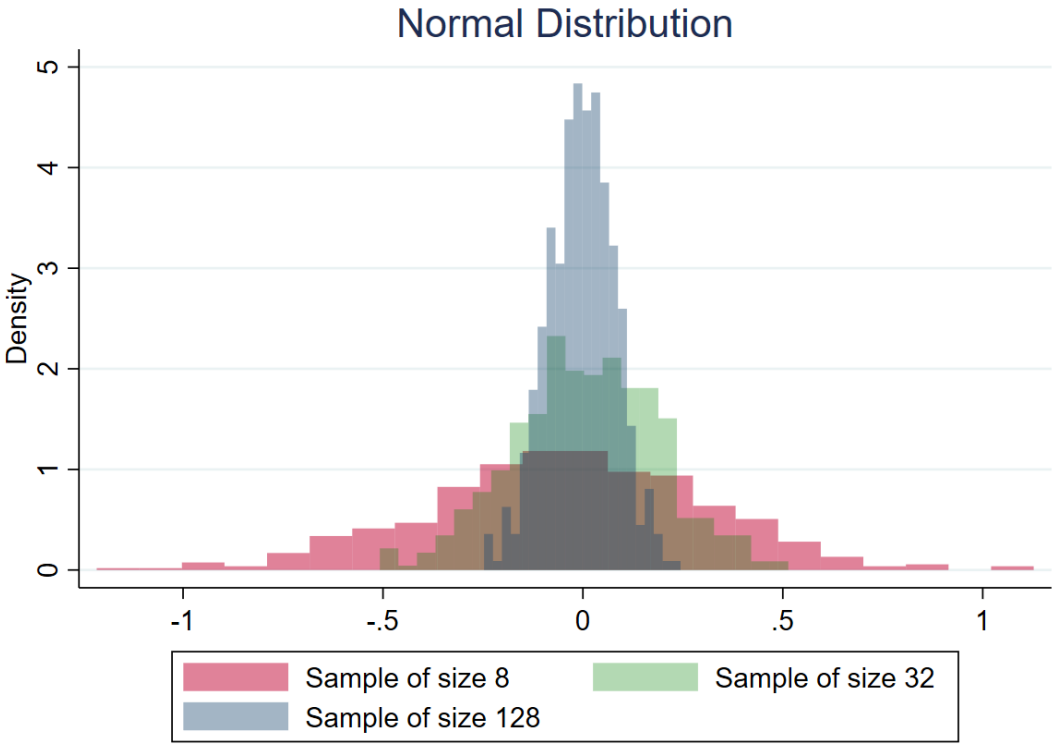
Hypothesis testing for weeks of insured unemployment: Employer experiment (EE)

$$T_n = \frac{\bar{Y}_t^{EE} - \bar{Y}_c^{EE}}{SE^{EE}} = \frac{17.7 - 18.3}{\sqrt{0.205^2 + 0.205^2}} \approx -\frac{0.6}{0.290} \approx -2.07$$

If we assume a 5% significance level, the corresponding one-tailed and two-tailed critical values will be -1.645 and 1.96 respectively. In the one-tailed test, we reject the null hypothesis that the experiment reduced the weeks of insured unemployment in the first unemployment spell as $T_n = -2.07 < -1.645$. In the two-tailed test, we also reject the null hypothesis at the 5% significance level since $|T_n| = 2.07 > 1.96$.

Part B: Problem 2a.





Part B: Problem 2b.

The summary statistics of the drawn distributions and the corresponding histograms confirm that as sample size increases, sample mean becomes more and more closer to the true population mean and the sampling variance decreases.

The first property is consistent with the law of large numbers (LLN), which implies convergence to the true population mean in probability for the sample mean as n increases. The second property is driven by the central limit theorem (CLT), according to which if we take a random sample of size n from a given distribution with mean μ and variance σ^2, the sampling mean will approximately have a normal distribution with mean μ and variance σ^2/n.

In our example, we can see that the sampling and theory-predicted standard deviations are almost the same, with the implication that the sampling variance, indeed, decreases at the rate predicted by the theory. Moreover, both LLN and CLT hold regardless of whether the underlying data is normal or not, which makes them powerful tools in statistical inference.

Mean and SD of sampling distributions and theory-predicted SDs

	Sampling mean	**Sampling SD**	**Theory-predicted SD**
Bernoulli (0.5)			
8	0.490	0.177	$0.5/\sqrt{8} = 0.177$
32	0.497	0.087	$0.5/\sqrt{32} = 0.088$
128	0.500	0.045	$0.5/\sqrt{128} = 0.044$
Standard normal (0,1)			
8	-0.039	0.351	$1/\sqrt{8} = 0.354$
32	0.009	0.182	$1/\sqrt{32} = 0.177$
128	-0.004	0.083	$1/\sqrt{128} = 0.088$

Part C: Problem 1

$$T_n = \frac{\bar{Y}_{HI}^{husbands} - \bar{Y}_{No\ HI}^{husbands}}{SE(\bar{Y}_{HI}^{husbands} - \bar{Y}_{No\ HI}^{husbands})} = \frac{0.31}{0.03} \approx 10.3$$

If we assume a 1% significance level, the corresponding two-tailed critical value will be 2.58. Since $|T_n|$ is greater than 2.58, we can reject the null hypothesis that husbands with and without health insurance have the same health status.

Part C: Problem 2a

Health and demographic characteristics of insured and uninsured husbands in the NHIS

	Some HI (1)	No HI (2)	Difference (3)
A. Health			
Health index	4.01 [0.93]	3.70 [1.01]	0.31 (0.03)
B. Other characteristics			
Nonwhite	0.16	0.17	-0.01 (0.01)
Age	43.98	41.26	2.71 (0.29)
Education	14.31	11.56	2.74 (0.10)
Family Size	3.50	3.98	-0.47 (0.05)
Employed	0.92	0.85	0.07 (0.01)
Family income	106467	45656	60810 (1356)
Sample size	8114	1281	

Standard deviations are in brackets; standard errors are reported in parentheses.

Part C: Problem 2b

Restricting the sample to employed college graduates reduces the difference in health index **(hlth)** between husbands with and without health insurance **(hi)** from 0.31 to 0.16. However, this difference is not statistically different from zero, as the corresponding t-statistic is 1.05 – less than the acceptable thresholds to reject the null hypothesis of zero health gap.

.

	(1)
	hlth
hi	0.160
	(1.05)
_cons	4.166***
	(27.61)
N	1125

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Part C: Problem 2c

The table below shows the regression results of health index **(hlth)** and health insurance status **(hi)** on being employed and having 16+ years of schooling **(emp_grad)**. In both specifications, the regression coefficients are statistically different from zero, which means that employed college graduates are generally healthier than the rest of the husbands in our sample and that the insured are more likely than the uninsured to be employed college graduates.

In previous section, we have shown that the health gap between the insured and uninsured disappears when we restrict the sample to employed college graduates. Combining this with the new findings, we can conclude that the difference in health between insured and uninsured NHIS husbands at least partly reflects the extra schooling of the insured. This might be because more educated people exercise more, apply for healthcare services if there is a need, smoke less, etc.

	(1)	(2)
	hlth	hi
emp_grad	0.409***	0.126***
	(14.10)	(19.57)
_cons	3.913***	0.847***
	(308.22)	(187.59)
N	9395	9395

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Statistics/Data analysis

```
*** Problem B2 *** Prepared by Gevorg Minasyan (February, 2023)

    pause on
    clear all
    set more off
    capture log close

*** Set to directory
    cd "D:\MIT DEDP\14.320\PS1"

    capture log using problem_B2.log, text replace
        set seed 12345

*** Random draws from Standard Normal and Bernoulli (0.5) distributions
          foreach i in 8 32 128 {
             set obs `i'
             forvalues j = 1(1)500 {
               generate n`i'_`j' = rnormal(0,1)
                   generate b`i'_`j' = rbinomial(1,0.5)
          }
          }


*** Converting matrices of sampling means into column variables

          order _all, first sequential

          foreach num in 8 32 128 {
              cap tabstat b`num'_1-b`num'_500, s(mean) save
              mat Bernoulli`num' = r(StatTotal)'
                  svmat Bernoulli`num', names(matcol)

                  cap tabstat n`num'_1-n`num'_500, s(mean) save
              mat Normal`num' = r(StatTotal)'
                  svmat Normal`num', names(matcol)
          }


*** Plotting the histograms of sampling means

twoway (histogram Bernoulli8Mean, color(cranberry%50) lcolor(none)) (histogram Bernoul
> li32Mean ///
     ,color(green%30) lcolor(none)) (histogram Bernoulli128Mean, color(navy%40) lcolor
> (none)) ///
     ,legend(order(1 "Sample of size 8" 2 "Sample of size 32" 3 "Sample of size 128"))
>  title("Bernoulli Distribution")

graph export "Bernoulli.png", replace

twoway (histogram Normal8Mean, color(cranberry%50) lcolor(none)) (histogram Normal32Me
> an ///
     ,color(green%30) lcolor(none)) (histogram Normal128Mean, color(navy%40) lcolor(no
> ne)) ///
         ,legend(order(1 "Sample of size 8" 2 "Sample of size 32" 3 "Sample of size 12
> 8")) title("Normal Distribution")

graph export "Normal.png", replace

tabstat Bernoulli8Mean Bernoulli32Mean Bernoulli128Mean Normal8Mean Normal32Mean Norma
> l128Mean ///
     ,statistics(mean sd) varwidth(50) columns(statistics) format(%8.3f)


capture log close
```

```
.         set seed 12345

.
. *** Random draws from Standard Normal and Bernoulli (0.5) distributions
.           foreach i in 8 32 128 {
  2.              set obs `i'
  3.              forvalues j = 1(1)500 {
  4.                generate n`i'_`j' = rnormal(0,1)
  5.                    generate b`i'_`j' = rbinomial(1,0.5)
  6.           }
  7.           }
Number of observations (_N) was 0, now 8.
Number of observations (_N) was 8, now 32.
Number of observations (_N) was 32, now 128.

.
.
. *** Converting matrices of sampling means into column variables
.
.           order _all, first sequential

.
.           foreach num in 8 32 128 {
  2.                cap tabstat b`num'_1-b`num'_500, s(mean) save
  3.                mat Bernoulli`num' = r(StatTotal)'
  4.                    svmat Bernoulli`num', names(matcol)
  5.
.                  cap tabstat n`num'_1-n`num'_500, s(mean) save
  6.                mat Normal`num' = r(StatTotal)'
  7.                    svmat Normal`num', names(matcol)
  8.           }
number of observations will be reset to 500
Press any key to continue, or Break to abort
Number of observations (_N) was 128, now 500.

.
.
. *** Plotting the histograms of sampling means
.
. twoway (histogram Bernoulli8Mean, color(cranberry%50) lcolor(none)) (histogram Bern
> oulli32Mean ///
>      ,color(green%30) lcolor(none)) (histogram Bernoulli128Mean, color(navy%40) lco
> lor(none)) ///
>      ,legend(order(1 "Sample of size 8" 2 "Sample of size 32" 3 "Sample of size 128
> ")) title("Bernoulli Distribution")

.
. graph export "Bernoulli.png", replace
file Bernoulli.png saved as PNG format

.
. twoway (histogram Normal8Mean, color(cranberry%50) lcolor(none)) (histogram Normal3
> 2Mean ///
>      ,color(green%30) lcolor(none)) (histogram Normal128Mean, color(navy%40) lcolor
> (none)) ///
>         ,legend(order(1 "Sample of size 8" 2 "Sample of size 32" 3 "Sample of size
>  128")) title("Normal Distribution")

.
```

```
. graph export "Normal.png", replace
file Normal.png saved as PNG format

.
. tabstat Bernoulli8Mean Bernoulli32Mean Bernoulli128Mean Normal8Mean Normal32Mean No
> rmal128Mean ///
>       ,statistics(mean sd) varwidth(50) columns(statistics) format(%8.3f)
(option varwidth() outside valid range 8..32; 32 assumed)

                         Variable |      Mean        SD
----------------------------------+--------------------
                  Bernoulli8Mean |     0.490     0.177
                 Bernoulli32Mean |     0.497     0.087
                Bernoulli128Mean |     0.500     0.045
                     Normal8Mean |    -0.039     0.351
                    Normal32Mean |     0.009     0.182
                   Normal128Mean |    -0.004     0.083
-------------------------------------------------------

.
.       capture log close
```

Statistics/Data analysis

```
*** Problem C2
*** Prepared by Gevorg Minasyan (February, 2023)

    pause on
    clear all
    set more off
    capture log close


*** Set to directory where NHIS2009_clean.dta is stored
    cd "D:\MIT DEDP\14.320\PS1"

    capture log using problem_C2.log, text replace

        use NHIS2009_clean, clear

*** Labels of the variables for Table 1.1
    label variable hlth "Health index 1-5"
    label variable nwhite "Nonwhite"
    label variable age "Age"
    label variable yedu "Education (years)"
    label variable famsize "Family size"
    label variable empl "Employed"
    label variable inc "Family income"
    label variable hi "Health insurance status"

*** PART I: Keep couples and select sample

*** Select non-missing HI respondents with one female in HH
*** Rrespondents are considered as having insurance if their spouse does
        keep if marradult == 1 & perweight != 0
        keep if hi != .
        by serial: egen numfem = total(fml)
        keep if numfem == 1
        drop numfem

*** MM T1.1 sample selection criteria
        gen T11 = ( age >= 26 & age <= 59 & marradult == 1 & adltempl >=1)
        keep if T11 == 1

*** Drop single-person HHs
        by serial: gen n = _N
        keep if n > 1

* PART II: Create different datasets for husbands and wives
        preserve
                keep if fml == 0
                save husbands.dta, replace
                sum hlth hi age marstat sex famsize relate racenew educ
        restore
                keep if fml == 1
                save wives.dta, replace
                sum hlth hi age marstat sex famsize relate racenew educ


*** a) Replicating the first 3 columns of Table 1.1
    use husbands.dta, replace

        matrix husbands  = J(15,3,.)
        matrix rownames husbands = "Health index" "se" "Nonwhite" "se" "Age" "se" "Edu
> cation" "se" "Family Size" "se" "Employed" "se" "Family income" "se" "Sample size"
        matrix colnames husbands = "Some HI" "No HI" "Difference"
```

```
*** HI with and without insurance for husbands

        qui sum hlth if hi == 1 [aw = perweight]
                mat husbands[1,1] = r(mean)
                mat husbands[2,1] = r(sd)

        qui sum hlth if hi == 0 [aw = perweight]
                mat husbands[1,2] = r(mean)
                mat husbands[2,2] = r(sd)

        reg hlth hi [aw=perweight], robust
                mat husbands[1,3] = _b[hi]
                mat husbands[2,3] = _se[hi]

*** Other characteristics of husbands with and without health insurance

    local i1 = 3
        local i2 = 4

        foreach var in nwhite age yedu famsize empl inc {

                local j = 1

                * Means
                        qui sum `var' if hi == 1 [aw = perweight]
                                mat husbands[`i1',`j'] = r(mean)
                                local ++ j

                        qui sum `var' if hi == 0 [aw = perweight]
                                mat husbands[`i1',`j'] = r(mean)
                                local ++ j

                * Differences and SEs
                        reg `var' hi [w = perweight], robust
                                mat husbands[`i1',`j'] = _b[hi]
                                mat husbands[`i2',`j'] = _se[hi]
                                local ++ j

                local i1 = `i1' + 2
                local i2 = `i2' + 2

                }



*** Adding sample size
        tab hi [aw=perweight], matcell(temp)
        mat husbands[`i1',2] = temp[1,1]
        mat husbands[`i1',1] = temp[2,1]


    matrix list husbands, format(%8.2f)


*** Output Table 1.1
    putexcel set Table11, modify
        putexcel A1 = matrix(husbands), names nformat(number_d2)


*** b) Restricting the sample to employed college graduates
```

```
    gen emp_grad =  empl == 1  & yedu > 16

        eststo: quietly reg hlth hi [aw=perweight] if emp_grad == 1, robust
    esttab

*** c) Comparison of employed college graduates with others

        eststo clear
        eststo: reg hlth emp_grad [aw=perweight], robust
        eststo: reg hi emp_grad [aw=perweight], robust
        esttab

capture log close
```

```
.
.         use NHIS2009_clean, clear

.
. *** Labels of the variables for Table 1.1
.     label variable hlth "Health index 1-5"

.     label variable nwhite "Nonwhite"

.     label variable age "Age"

.     label variable yedu "Education (years)"

.     label variable famsize "Family size"

.     label variable empl "Employed"

.     label variable inc "Family income"

.     label variable hi "Health insurance status"

.
. *** PART I: Keep couples and select sample
.
. *** Select non-missing HI respondents with one female in HH
. *** Rrespondents are considered as having insurance if their spouse does
.         keep if marradult == 1 & perweight != 0
(50,662 observations deleted)

.         keep if hi != .
(0 observations deleted)

.         by serial: egen numfem = total(fml)

.         keep if numfem == 1
(49 observations deleted)

.         drop numfem

.
. *** MM T1.1 sample selection criteria
.         gen T11 = ( age >= 26 & age <= 59 & marradult == 1 & adltempl >=1)

.         keep if T11 == 1
(9,657 observations deleted)

.
. *** Drop single-person HHs
.         by serial: gen n = _N

.         keep if n > 1
(1,476 observations deleted)

.
. * PART II: Create different datasets for husbands and wives
.         preserve

.                 keep if fml == 0
(9,395 observations deleted)
```

```
.                 save husbands.dta, replace
file husbands.dta saved

.                 sum hlth hi age marstat sex famsize relate racenew educ

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        hlth |      9,395    3.931666    .9524645          1          5
          hi |      9,395    .8372539    .3691535          0          1
         age |      9,395    43.69207    8.641379         26         59
     marstat |      9,395          10           0         10         10
         sex |      9,395           1           0          1          1
-------------+---------------------------------------------------------
     famsize |      9,395    3.633209    1.369862          2         18
      relate |      9,395    14.70676    4.991659         10         20
     racenew |      9,395    13.89888    9.165081         10         50
        educ |      9,395    15.95679    3.565478          1         22

.         restore

.                 keep if fml == 1
(9,395 observations deleted)

.                 save wives.dta, replace
file wives.dta saved

.                 sum hlth hi age marstat sex famsize relate racenew educ

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        hlth |      9,395    3.933262     .952857          1          5
          hi |      9,395    .8461948    .3607811          0          1
         age |      9,395    41.74593    8.647144         26         59
     marstat |      9,395          10           0         10         10
         sex |      9,395           2           0          2          2
-------------+---------------------------------------------------------
     famsize |      9,395    3.633209    1.369862          2         18
      relate |      9,395    15.29324    4.991659         10         20
     racenew |      9,395    14.06493    9.404566         10         50
        educ |      9,395    16.16179    3.447716          1         22

.
.
. *** a) Replicating the first 3 columns of Table 1.1
.     use husbands.dta, replace

.
.         matrix husbands  = J(15,3,.)

.         matrix rownames husbands = "Health index" "se" "Nonwhite" "se" "Age" "se" "
> Education" "se" "Family Size" "se" "Employed" "se" "Family income" "se" "Sample siz
> e"

.         matrix colnames husbands = "Some HI" "No HI" "Difference"

.
.
. *** HI with and without insurance for husbands
.
.         qui sum hlth if hi == 1 [aw = perweight]
```

```
.                 mat husbands[1,1] = r(mean)

.                 mat husbands[2,1] = r(sd)

.
.         qui sum hlth if hi == 0 [aw = perweight]

.                 mat husbands[1,2] = r(mean)

.                 mat husbands[2,2] = r(sd)

.
.         reg hlth hi [aw=perweight], robust
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =      84.68
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0129
                                                Root MSE          =      .9406

------------------------------------------------------------------------------
             |               Robust
        hlth | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |   .3132452   .0340396     9.20   0.000     .2465202    .3799702
       _cons |   3.695654   .0316859   116.63   0.000     3.633543    3.757765
------------------------------------------------------------------------------

.                 mat husbands[1,3] = _b[hi]

.                 mat husbands[2,3] = _se[hi]

.
. *** Other characteristics of husbands with and without health insurance
.
.     local i1 = 3

.         local i2 = 4

.
.         foreach var in nwhite age yedu famsize empl inc {
  2.
.                 local j = 1
  3.
.                 * Means
.                         qui sum `var' if hi == 1 [aw = perweight]
  4.                                 mat husbands[`i1',`j'] = r(mean)
  5.                                 local ++ j
  6.
.                         qui sum `var' if hi == 0 [aw = perweight]
  7.                                 mat husbands[`i1',`j'] = r(mean)
  8.                                 local ++ j
  9.
.                 * Differences and SEs
.                         reg `var' hi [w = perweight], robust
 10.                                 mat husbands[`i1',`j'] = _b[hi]
 11.                                 mat husbands[`i2',`j'] = _se[hi]
 12.                                 local ++ j
 13.
```

```
.                 local i1 = `i1' + 2
 14.                  local i2 = `i2' + 2
 15.
.                 }
(analytic weights assumed)
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =       1.01
                                                Prob > F          =     0.3144
                                                R-squared         =     0.0001
                                                Root MSE          =     .36602

------------------------------------------------------------------------------
             |               Robust
      nwhite | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |  -.0115948   .0115249    -1.01   0.314     -.034186    .0109965
       _cons |   .1693667   .0106274    15.94   0.000     .1485347    .1901986
------------------------------------------------------------------------------
(analytic weights assumed)
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =      86.70
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0114
                                                Root MSE          =      8.661

------------------------------------------------------------------------------
             |               Robust
         age | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |   2.713743   .2914405     9.31   0.000     2.142457    3.285029
       _cons |   41.26318   .2671254   154.47   0.000     40.73956    41.78681
------------------------------------------------------------------------------
(analytic weights assumed)
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =     729.63
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1129
                                                Root MSE          =       2.64

------------------------------------------------------------------------------
             |               Robust
        yedu | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |   2.743283   .1015592    27.01   0.000     2.544205    2.942361
       _cons |   11.56244   .0962655   120.11   0.000     11.37374    11.75115
------------------------------------------------------------------------------
(analytic weights assumed)
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =      81.36
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0147
                                                Root MSE          =     1.3262

------------------------------------------------------------------------------
             |               Robust
     famsize | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |  -.4721901   .0523507    -9.02   0.000    -.5748089   -.3695713
       _cons |   3.976924   .0494789    80.38   0.000     3.879934    4.073913
------------------------------------------------------------------------------
(analytic weights assumed)
(sum of wgt is 34,118,563)
```

```
Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =      40.01
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0084
                                                Root MSE          =     .27911

------------------------------------------------------------------------------
             |               Robust
        empl | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |   .0748071   .0118272     6.33   0.000     .0516233     .097991
       _cons |    .849466   .0112809    75.30   0.000     .8273529    .8715791
------------------------------------------------------------------------------
(analytic weights assumed)
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =    2011.74
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1380
                                                Root MSE          =      52163

------------------------------------------------------------------------------
             |               Robust
         inc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          hi |   60810.44   1355.789    44.85   0.000      58152.8    63468.08
       _cons |   45656.25   1149.919    39.70   0.000     43402.16    47910.34
------------------------------------------------------------------------------

.
.
.
. *** Adding sample size
.         tab hi [aw=perweight], matcell(temp)

    Health |
  insurance |
     status |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 | 1,281.4929       13.64       13.64
          1 | 8,113.5071       86.36      100.00
------------+-----------------------------------
      Total |      9,395      100.00

.         mat husbands[`i1',2] = temp[1,1]

.         mat husbands[`i1',1] = temp[2,1]

.
.
.     matrix list husbands, format(%8.2f)

husbands[15,3]
                Some HI       No HI  Difference
Health index       4.01        3.70        0.31
          se       0.93        1.01        0.03
    Nonwhite       0.16        0.17       -0.01
          se          .           .        0.01
         Age      43.98       41.26        2.71
          se          .           .        0.29
   Education      14.31       11.56        2.74
          se          .           .        0.10
 Family Size       3.50        3.98       -0.47
          se          .           .        0.05
    Employed       0.92        0.85        0.07
          se          .           .        0.01
Family inc~e    1.1e+05    45656.25    60810.44
          se          .           .     1355.79
 Sample size    8113.51     1281.49           .
```

```
.
.
. *** Output Table 1.1
.     putexcel set Table11, modify

.         putexcel A1 = matrix(husbands), names nformat(number_d2)
file Table11.xlsx saved

.
.
. *** b) Restricting the sample to employed college graduates
.
.     gen emp_grad =  empl == 1  & yedu > 16

.
.         eststo: quietly reg hlth hi [aw=perweight] if emp_grad == 1, robust
(est1 stored)

.     esttab

----------------------------
                      (1)
                     hlth
----------------------------
hi                   0.160
                    (1.05)

_cons                4.166***
                   (27.61)
----------------------------
N                     1125
----------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

.
. *** c) Comparison of employed college graduates with others
.
.         eststo clear

.         eststo: reg hlth emp_grad [aw=perweight], robust
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =     198.85
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0211
                                                Root MSE          =      .9367

------------------------------------------------------------------------------
             |               Robust
        hlth | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    emp_grad |   .4085993   .0289761    14.10   0.000     .3517999    .4653987
       _cons |   3.913048   .0126957   308.22   0.000     3.888161    3.937934
------------------------------------------------------------------------------
(est1 stored)

.         eststo: reg hi emp_grad [aw=perweight], robust
(sum of wgt is 34,118,563)

Linear regression                               Number of obs     =      9,395
                                                F(1, 9393)        =     383.17
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0152
                                                Root MSE          =     .34064
```

```
------------------------------------------------------------------------------
             |               Robust
          hi | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    emp_grad |   .1256117   .0064171    19.57   0.000     .1130329    .1381906
       _cons |   .8472668   .0045167   187.59   0.000     .8384131    .8561205
------------------------------------------------------------------------------
(est2 stored)

.         esttab

--------------------------------------------
                      (1)             (2)
                     hlth              hi
--------------------------------------------
emp_grad            0.409***        0.126***
                   (14.10)         (19.57)

_cons               3.913***        0.847***
                  (308.22)        (187.59)
--------------------------------------------
N                     9395            9395
--------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

.
. capture log close
```