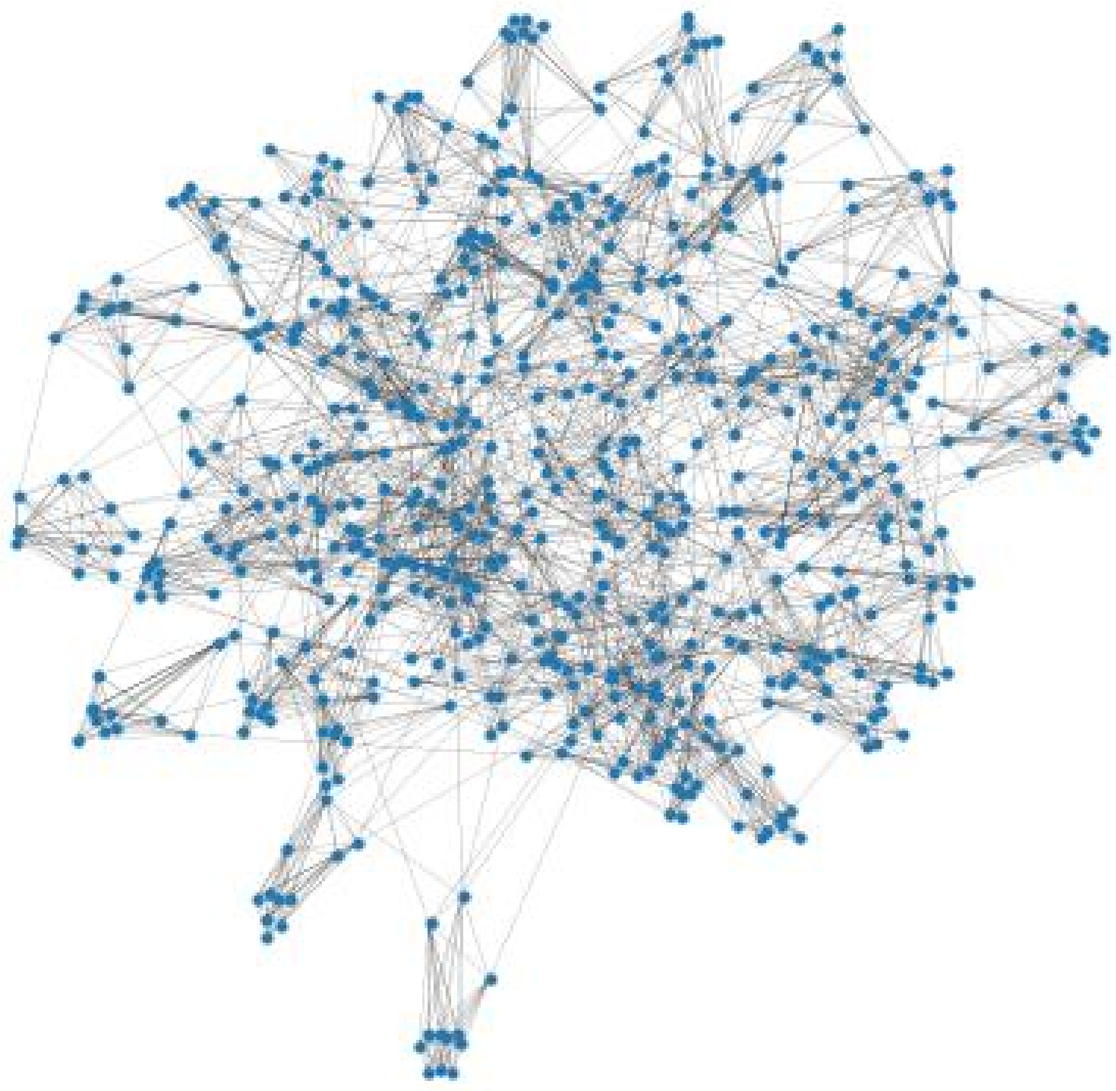
RELAXE CAVEMAN GRAPH



a) (3 pts) First, let's implement the three sampling schemes. **Program** three functions that perform:

1. **Bernoulli sampling of nodes**: iterate over nodes, and sample each one with probability $p$. We observe an edge if and only if we have sampled its two constituting nodes.
2. **Bernoulli sampling of edges**: iterate over edges, and sample each one with probability $p$. We observe a node if and only if we have sampled at least one of its edges.
3. **Star sampling**: iterate over nodes, and sample each one with probability $p$. If you have directly sampled a node, you also observe all of its neighbors (a real-life example would be a dataset obtained by crawling through friendship lists of randomly selected users in a social networking website). Note: here we will have nodes that are sampled a) directly with probability $p$ and b) indirectly via sampling a neighbor. It is useful to keep a list of nodes you sampled directly.

**Obtain samples** of the network using the three sampling schemes with probability $p = 0.28$. Then, use either the template code or your own to obtain empirical estimates of the number of triangles, the number of two-stars, and transitivity. **Report** your results on a table where the columns represent:

- sampled number of triangles,
- sampled number of two-stars,
- transitivity in sampled network,
- fraction of sampled triangles over triangles in original network,
- fraction of sampled two-stars over two-stars in original network,

and the rows represent the different sampling schemes (plus an extra row with the values of the original network). **Answer** the following questions:

- How do sampling schemes compare in the fraction of triangles/two-stars they preseve?
- Do sampling schemes affect two-stars/triangles in the same way?
- Transitivity via node sampling should be similar to the real value; what could be the reason for this?

| samp. | triangles | two-stars | transit. | triang.frac. | two-st.frac. |
|---|---|---|---|---|---|
| node | 468 | 657 | 0.7123 | 0.0181 | 0.0179 |
| edge | 657 | 3109 | 0.2113 | 0.0255 | 0.0847 |
| star | 5649 | 10923 | 0.5172 | 0.2190 | 0.2977 |
| orig. | 25800 | 36697 | 0.7031 | 1.0000 | 1.0000 |

- If we compare the fraction of preserved triangles we can see that with the node and edge samplings we obtain similar values, while with the star sampling is higher. On the other hand, the fraction of preserved two stars is affected differently: the value with the star sampling is ~3.5 bigger than the one with edge sampling, and this last one is ~4.5 bigger than the value obtained with the node sampling. So, between the three methods, the star sampling is the best.

- Observing the results we can conclude that node and edge sampling schemes affect in a "similar" way the number of triangles, but not the number of two-stars. And in both cases the effect is different with the star sampling scheme.

- Compare to the original network we can observe that the transitivity via node sample is similar to the real value, this happens because sample all nodes with the same probability will tend to preserve the same* fraction of triangles and two-stars. This result is showed in the next exercises

b) (2 pts, pen and paper) Different sampling schemes alter the observed number of structures on the sampled networks. Luckily, knowing the sampling probability $p$, we can estimate how much these numbers vary. As an example, for Bernoulli sampling of **edges**, the probability of sampling a two-star from the original network is $p^e_{\angle} = p^2$ (we need to observe two edges), while a triangle is sampled with probability $p^e_{\triangle} = p^3$ (we need to observe three edges). **Derive and explain how to obtain**:

i) the probabilities of sampling a two-star and a triangle using Bernoulli sampling of **nodes** ($p^n_{\angle}$ and $p^n_{\triangle}$); and

- $p^n_{\angle} = p^3$



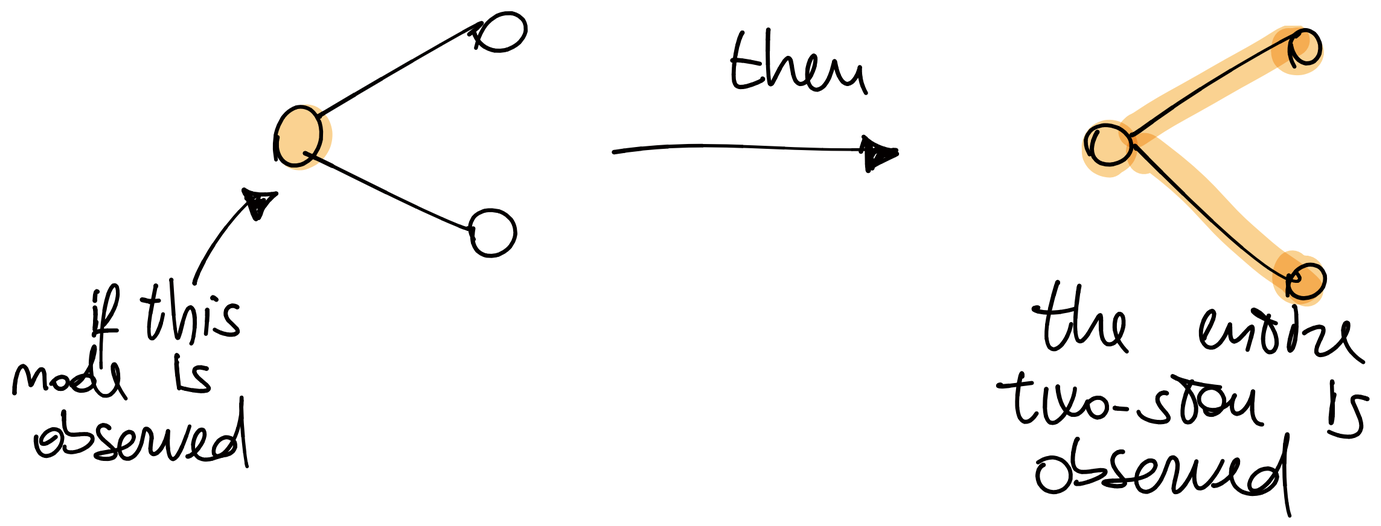
- $p^n_{\triangle} = p^3$



In both cases, for a star and for a triangle, you must observe all the three nodes, and considering that the prob of observing a node is $p$, the prob of observing three nodes is $p^3$.

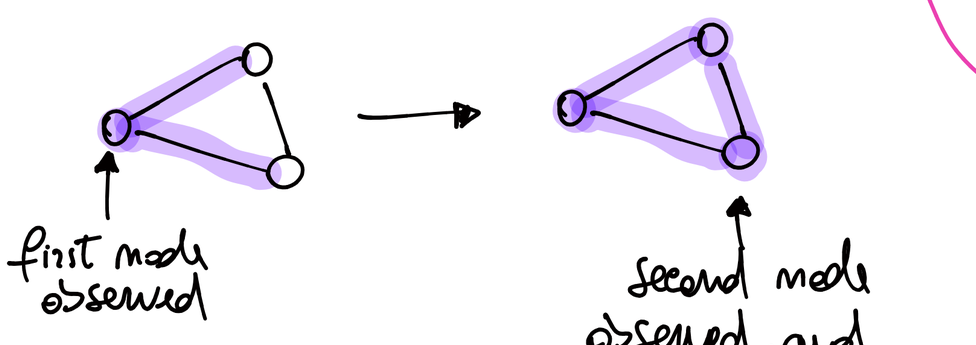
ii) the probabilities of sampling a two-star and a triangle using **star sampling** ($p^s_{\angle}$ and $p^s_{\triangle}$).

- $p^s_{\angle} = p$

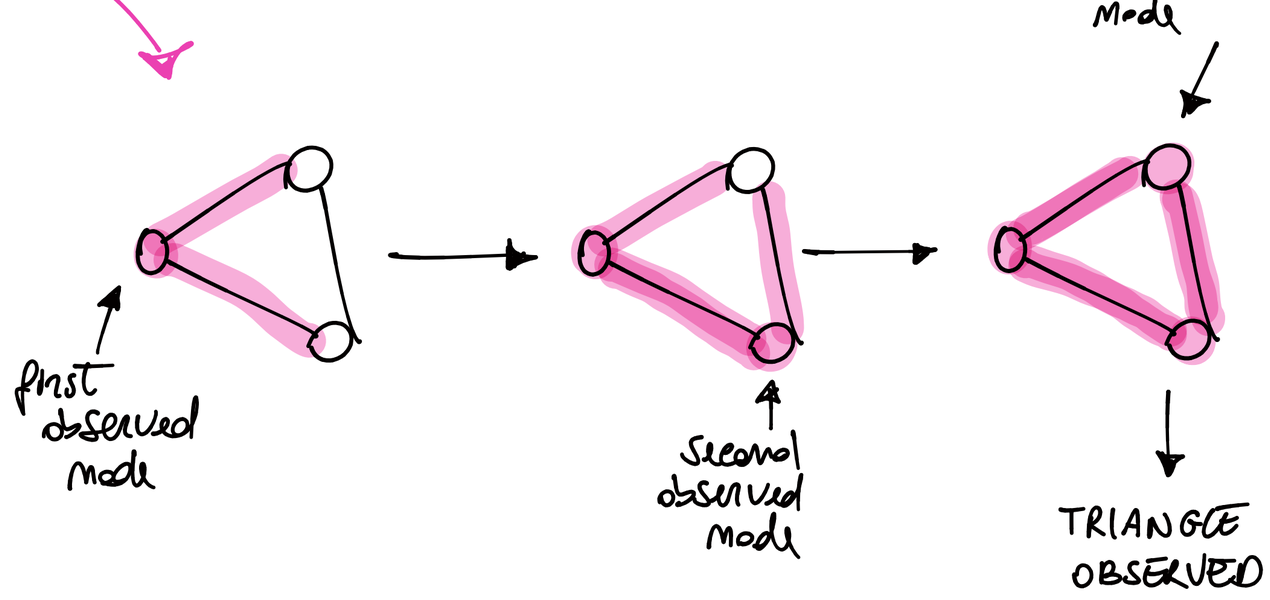



- $p^s_{\triangle} = 3 \cdot p^2 \cdot (1-p) + p^3$

since we need to observe at least 2 directed nodes, so we have 3 ways of observing 2 directed nodes, plus the probability of observing 3 directed nodes




(You have 3 possible configurations, if you change the first observed node)




c) (2 pt, pen and paper) The Horvitz-Thompson (HT) estimator is a simple way of correcting for the bias induced by sampling. Let $\hat{\tau}$ be the empirical count of a structure found in a sampled network, such as your results from a). If $p_\tau$ is the probability of observing these structures, then the HT estimator for the total counts is simply

$$\hat{\tau}^{HT} = \frac{1}{p_\tau}\hat{\tau}. \tag{2}$$

**Explain** why the HT estimator corrects for sampling bias. **Write and simplify** the HT estimators for the number of two-stars, triangles and transitivity for the three sampling schemes we explored before.

The Horvitz-Thompson (HT) estimator corrects for sampling bias by assigning each sampled unit a weight that accounts for its probability of being selected in the sample. This weighting ensures that the estimator provides an unbiased estimate of the population parameter, even when the sampling process is not simple random sampling. It basically adjusts for the fact that some units in the population are more or less likely to be included in the sample, making the estimator more representative of the entire population

BERNOULLI SAMPLING OF NODES

- $\hat{\tau}_A^{HT,n} = \frac{\hat{\tau}_A^n}{p_{\tau\Delta}} = \frac{\hat{\tau}_\Delta^n}{p^3}$
- $\hat{\tau}_L^{HT,n} = \frac{\hat{\tau}_L^n}{p_{\tau L}} = \frac{\hat{\tau}_L^n}{p^3}$
- $\hat{\tau}_C^{HT,n} = \frac{\hat{\tau}_\Delta^{HT,n}}{\hat{\tau}_L^{HT,n}} = \frac{\frac{\hat{\tau}_\Delta^n}{p^3}}{\frac{\hat{\tau}_L^n}{p^3}} = \frac{\hat{\tau}_\Delta^n}{\hat{\tau}_L^n}$

---

BERNOULLI SAMPLING OF EDGES

- $\hat{\tau}_A^{HT,e} = \frac{\hat{\tau}_\Delta^e}{p_{\tau\Delta}} = \frac{\hat{\tau}_\Delta^e}{p^3}$
- $\hat{\tau}_L^{HT,e} = \frac{\hat{\tau}_L^e}{p_{\tau L}} = \frac{\hat{\tau}_L^e}{p^2}$
- $\hat{\tau}_C^{HT,e} = \frac{\hat{\tau}_\Delta^{HT,e}}{\hat{\tau}_L^{HT,e}} = \frac{\frac{\hat{\tau}_\Delta^e}{p^3}}{\frac{\hat{\tau}_L^e}{p^2}} = \frac{\hat{\tau}_\Delta^e}{p \cdot \hat{\tau}_L^e}$

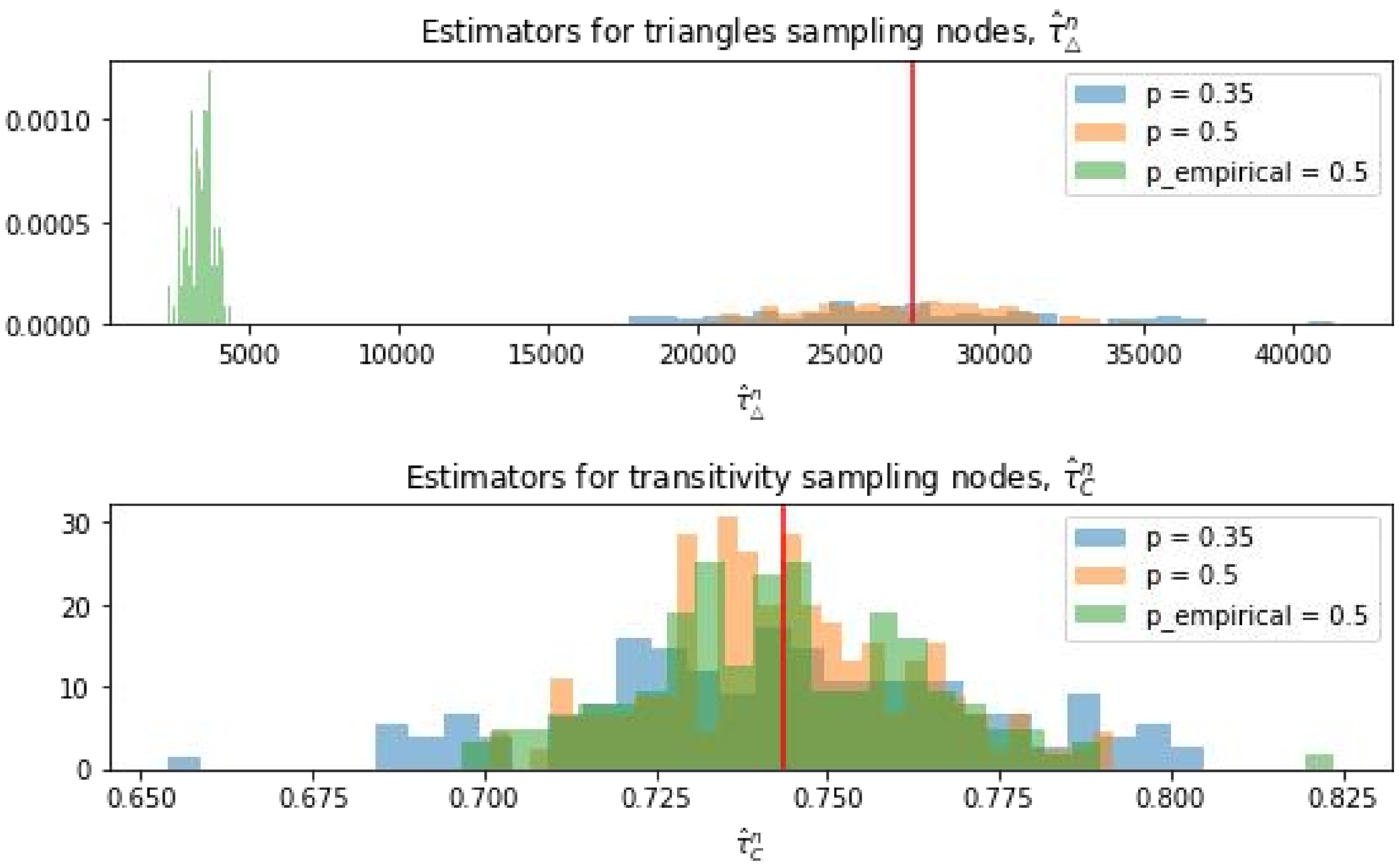
---

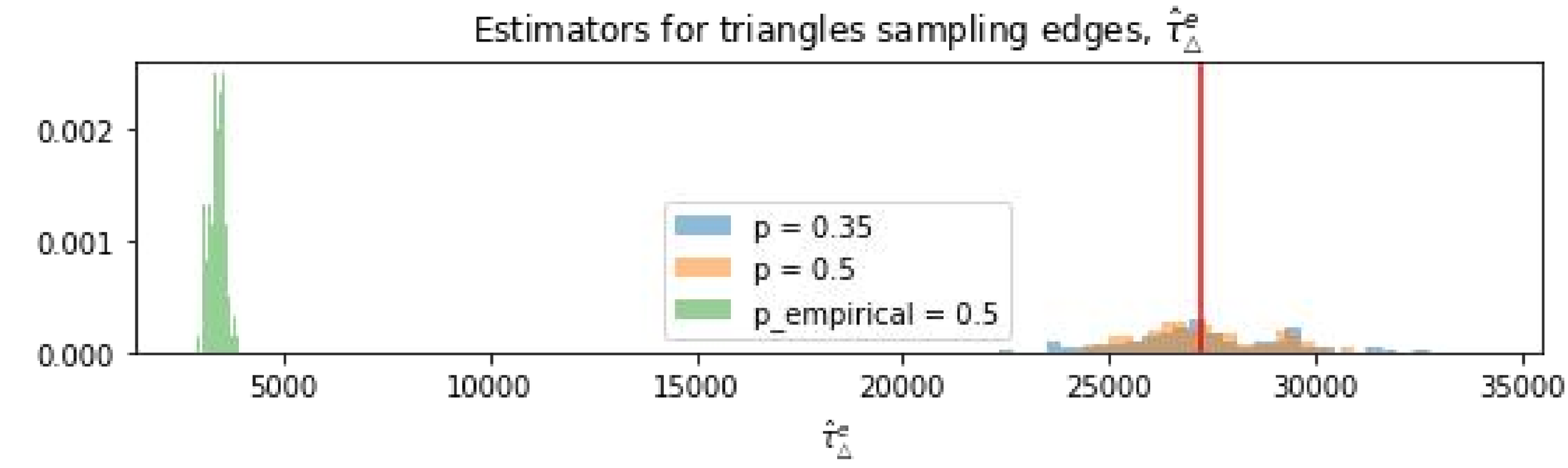
STAR SAMPLING

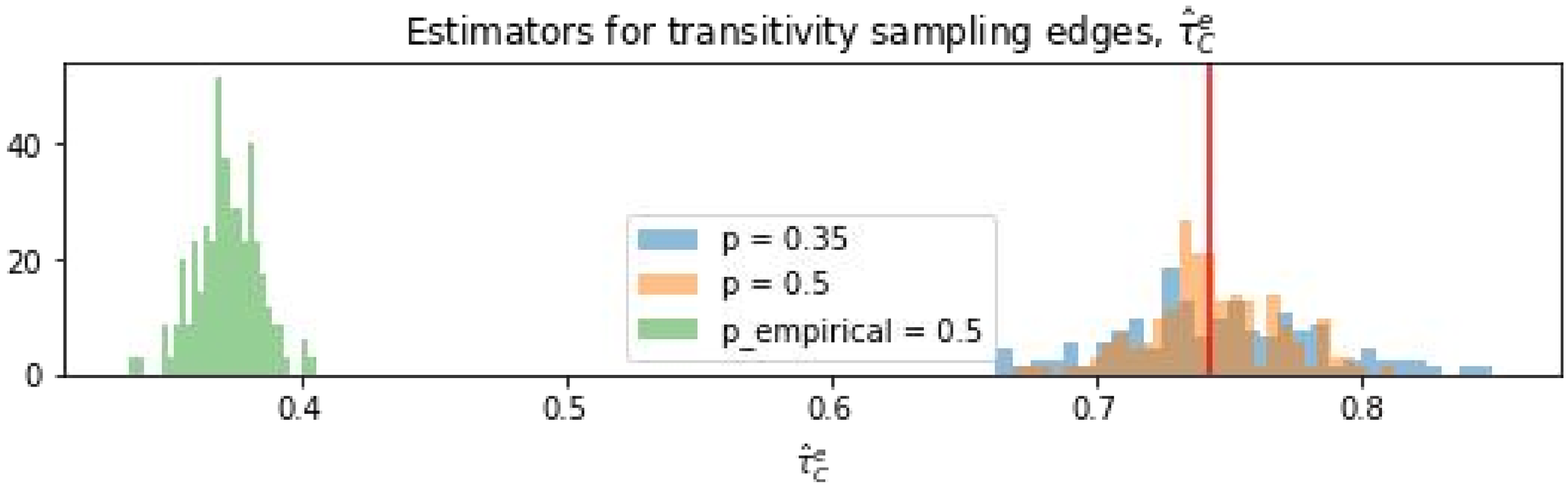
- $\hat{\tau}_A^{HT,s} = \frac{\hat{\tau}_\Delta^s}{p_{\tau\Delta}} = \frac{\hat{\tau}_\Delta^s}{3p^2(1-p)+p^3}$
- $\hat{\tau}_L^{HT,s} = \frac{\hat{\tau}_L^s}{p_{\tau L}} = \frac{\hat{\tau}_L^s}{p}$
- $\hat{\tau}_C^{HT,s} = \frac{\hat{\tau}_\Delta^{HT,s}}{\hat{\tau}_L^{HT,s}} = \frac{\frac{\hat{\tau}_\Delta^s}{3p^2(1-p)+p^3}}{\frac{\hat{\tau}_p^s}{p}} = \frac{\hat{\tau}_\Delta^s}{\hat{\tau}_p^s (3-2p)p}$

d) (3 pts) **Implement** the HT estimator for the three sampling schemes. Given two selection probabilities $p$, we will sample at least $n = 150$ times to obtain distributions of some of our HT estimators, and compare it with the measurements from the sampled networks. For each $p = 0.35, 0.5$, and for each sampling scheme, obtain $n = 150$ samples from the original network, and **calculate** the HT estimator for the number of triangles and for transitivity. For $p = 0.5$ include also the empirical estimator (the counts without the HT correction), and for all plots include a vertical line depicting the value of the original network. As a summary, you will **report your results in six plots**:

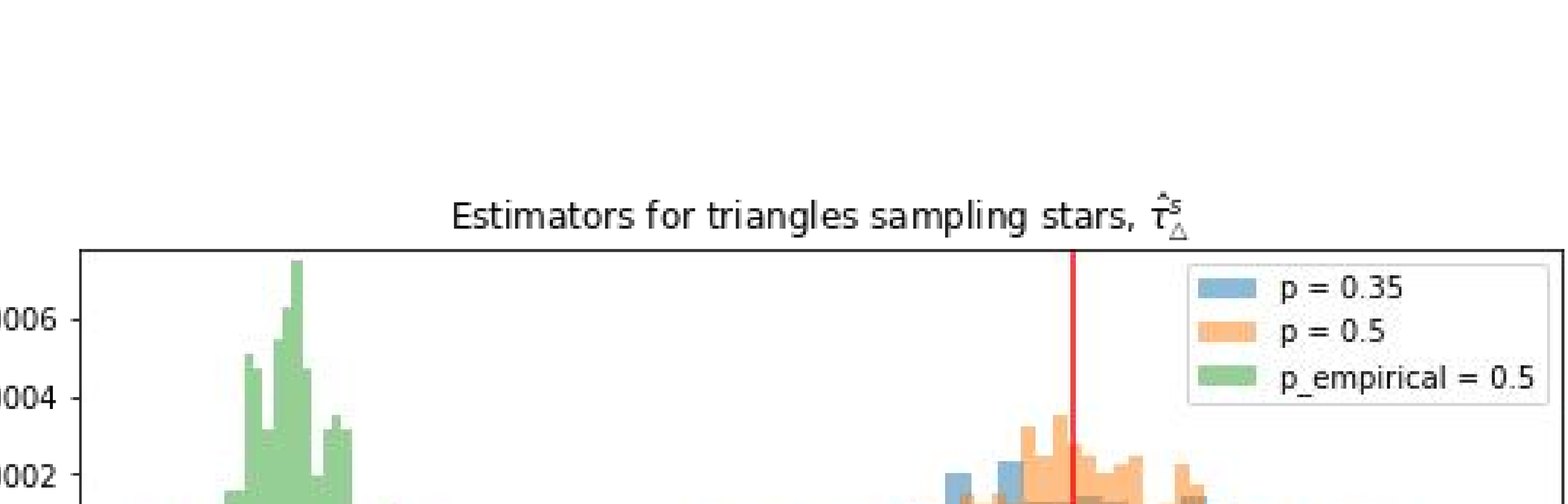
- Histograms of estimator $\hat{\tau}_{\triangle}^{HT,n}$ for $p = 0.35, 0.5$ and $\hat{\tau}_{\triangle}$ (empirical counts) for $p = 0.5$ and true value of $\tau_{\triangle}$.
- Histograms of estimator $\hat{\tau}_{C}^{HT,n}$ for $p = 0.35, 0.5$ and $\hat{\tau}_{C}$ (empirical value) for $p = 0.5$ and true value of $C$.
- Histograms of estimator $\hat{\tau}_{\triangle}^{HT,e}$ for $p = 0.35, 0.5$ and $\hat{\tau}_{\triangle}$ (empirical counts) for $p = 0.5$ and true value of $\tau_{\triangle}$.
- Histograms of estimator $\hat{\tau}_{C}^{HT,e}$ for $p = 0.35, 0.5$ and $\hat{\tau}_{C}$ (empirical value) for $p = 0.5$ and true value of $C$.
- Histograms of estimator $\hat{\tau}_{\triangle}^{HT,s}$ for $p = 0.35, 0.5$ and $\hat{\tau}_{\triangle}$ (empirical counts) for $p = 0.5$ and true value of $\tau_{\triangle}$.
- Histograms of estimator $\hat{\tau}_{C}^{HT,s}$ for $p = 0.35, 0.5$ and $\hat{\tau}^{C}$ (empirical value) for $p = 0.5$ and true value of $C$.




Estimators for triangles sampling edges, $\hat{\tau}^e_\triangle$
0.002
0.001
0.000
5000
10000
15000
20000
25000
30000
35000
p = 0.35
p = 0.5
p_empirical = 0.5
$\hat{\tau}^e_\triangle$


Estimators for transitivity sampling edges, $\hat{\tau}^e_C$
40
20
0
0.4
0.5
0.6
0.7
0.8
p = 0.35
p = 0.5
p_empirical = 0.5
$\hat{\tau}^e_C$


Estimators for triangles sampling stars, $\hat{\tau}^s_\triangle$
0.0006
0.0004
0.0002
0.0000
10000
15000
20000
25000
30000
35000
p = 0.35
p = 0.5
p_empirical = 0.5
$\hat{\tau}^s_\triangle$


Estimators for transitivity sampling stars, $\hat{\tau}^s_C$
30
20
10
0
0.675
0.700
0.725
0.750
0.775
0.800
0.825
p = 0.35
p = 0.5
p_empirical = 0.5
$\hat{\tau}^s_C$

In your plots, the distribution of HT estimators should lie around the real value. **Answer** the following questions:

- What is the effect of the sampling probability $p$ on your estimators?
- How do the HT distributions differ between sampling schemes?
- In the last plot, the empirical estimators (without the HT correction) should be centered around the true value; why is this the case?

When the sampling probability is increased, the weights assigned to sampled units become smaller, and conversely, if the probability is decreased, the weights become larger. I have noticed that with a higher probability, the distribution of values is more spread out around the true value compared to the case with lower probability. This phenomenon occurs because when the probability is increased, the estimator gives more weight to units that are less likely to be in the sample. As a result, the estimate becomes more sensitive to individual variations within the population, leading to a wider spread of values. In contrast, with lower probabilities, the estimator relies more on units with higher inclusion probabilities, which can result in a narrower distribution of values as it doesn't capture as much diversity within the population.

In the case of estimators for triangles/transitivity using sampled nodes, as well as in the case of estimators for triangles/transitivity using sampled edges, and in the case of the estimator for triangles using sampled stars, these are much more precise (i.e., centered around the true value) compared to the empirical estimator. It's also possible to observe that node sampling is not the best method for the number of triangles, and the results obtained with edge sampling are not satisfactory either. On the other hand, the star sampling method performs significantly better. Regarding transitivity, the estimates are quite satisfactory in all three cases, which is in line with the explanation provided in part a) of the exercise.

In the last plot, the empirical estimators (without HT correction) are expected to be centered around the true value primarily because a sampling probability of 0.5 was utilized. This choice corresponds to employing a simple random sampling method with equal probabilities for all units in the population. Under these conditions, the empirical estimator tends to be unbiased. The 0.5 sampling probability indicates that the sampled stars are representative of the overall population, which enhances the likelihood of the empirical estimator providing a centered and accurate estimate of the true value.