STA238 Notes

Probability Without Proofs II

https://github.com/ICPRplshelp/

Last updated January 11, 2023

1 Sampling Distribution and the Central Limit Theorem

IDEA – The sampling distribution is meant to give us a probability distribution for the sample mean and variance (being called statistic) if we took random samples out of a random variable. Because there is a chance we could take a sample and somehow end up with a sample mean being very far from μ (but it occurs very rarely – and the sampling distribution can tell us how likely that is to happen).

- We're learning about a method that will likely never be used in practice because of how difficult it is to get a random sample, unless our group is small.
- The main goal is that as long as our sample is completely random, we can guess the theoretical mean and variance, and ALSO calculate how confident we should be with our guess (that is, confidence intervals).

An **EXPERIMENTAL UNIT** is someone or something which we might collect data from.

A **POPULATION** is the set of ALL units we're interested in.

A **VARIABLE** is a characteristic or property of an individual unit from the population. Each person (or thing) has a property, right?

If we look at a population of people and Y_i represents the age of the ith person, we may have this space:

$$\boxed{Y_1, Y_2, \ldots, Y_{10}, \ldots, Y_N} \leftarrow N = 50000$$

If we have measurements for EVERYONE, then we can calculate the population mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

And the variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \mu)^2$$

Not feasible to get everyone. Right? Let's collect a random sample (no, we will never feasibly get that) of 200. With our sample, we can measure their age:

$$\boxed{y_1, y_2, \ldots, y_5, \ldots, y_n} \leftarrow n = 200$$

Usually, μ and σ, the population parameters, are unknown and are too difficult to measure, and that value can fluctuate. Good thing is that we can estimate close to that parameter.

After collecting the sample, we can measure some statistic(s): a function of the sample observation

- For example, the sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- And the sample standard deviation: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2$
  - The denominator contains a $n-1$ to get an unbiased estimator for σ^2.

1.1 Linking Population and Samples

Population:

$$\boxed{\mu,\ \sigma^2}^N \quad \text{UNKNOWN}$$
$$\boxed{\bar{y},\ s^2}^n \quad \text{KNOWN}$$

Sampling distribution helps bridge the gap between the unknown and the known.

We assume that $\bar{y}$ and s^2 are random variables. What are the sampling distributions for them? For the **samples:**

$$\begin{aligned} y_1, y_2, \ldots, y_n &\Leftarrow (\mu, \sigma) \\ E(y_i) &= \mu \\ V(y_i) &= \sigma^2 \\ i &= 1, 2 \ldots, n \end{aligned}$$

All the samples are identically distributed.

For the **statistic:**

$$\begin{aligned} E(\bar{y}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n} \cdot \mu \cdot n = \mu \\ V(\bar{y}) &= V\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(y_i) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n} \\ SD(\bar{y}) &= \sqrt{V(\bar{y})} = \frac{\sigma}{\sqrt{n}} \quad \text{standard error} \end{aligned}$$

So, IF $Y \sim (\mu, \sigma)$, THEN $\bar{y} \sim ?\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$?

CASE 1. σ is known, AND population is normal.

- Then, $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

CASE 2. σ is known, AND $n \geq 30$ (n is large).

- Then, by CLT, $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, irrespective of y_i's distribution.

CASE 3. σ is unknown, AND $n \geq 30$.

- $Z = \frac{(\bar{y}-\mu)}{\frac{s}{\sqrt{n}}} \sim N(0,\ 1)$
- Questions like $P(\bar{y} \leq c \cdot s)$ would be feasible to solve in this manner, but not necessarily $P(\bar{y} \leq c)$.

CASE 4. σ is unknown, population is normal, and $n < 30$.

- $T = \frac{\bar{y}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1 \leftarrow \text{degrees of freedom}}$
  - The $n-1$ stands for degrees of freedom

If you know these four cases, it's easier to make statistical inferences.

1.2 Sample Proportion

A **binary variable** takes only two outcomes, such as tossing a coin. When tossing a coin, we're able to get success or failure:

$$y_i = 1 \text{ if success otherwise } 0$$

Success will be denoted as S and failure will be denoted as F. Then, $Y \sim \text{Ber}(p)$, where p is the probability of success.

$$P(S) = p \quad P(F) = 1 - p$$

Hence $P(Y=y) = p^y(1-p)^{1-y}$. Then, for a Bernoulli distribution:

$$\mu = E(Y) = p$$
$$\sigma^2 = V(Y) = p(1-p)$$

Definition: Consider sample proportions:

$$\underset{\text{sample proportion}}{\widehat{p}} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

The total number of successes divided by sample size.

The sampling distribution for sample proportions:

$$E(\widehat{p}) = E(\bar{y}) = \mu = p$$
$$\Rightarrow \boxed{E(\widehat{p}) = p}$$
$$V(\widehat{p}) = V(\bar{y}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$$
$$\Rightarrow \boxed{V(\widehat{p}) = \frac{p(1-p)}{n}}$$
$$SD(\widehat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

CONDITION FOR NORMAL APPROXIMATION:

$(np \geq 10 \text{ and } n(1-p) \geq 10) \Rightarrow \widehat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

- Expected number of successes and expected number of failures both are ≥ 10.

1.3 T-Distribution

With condition in which we don't know σ, population is normal, $n < 30$:

$$T = \frac{\bar{y} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

You can find out how to solve T-distribution-type questions in the STA237 notes document. In practice, you'll always be calculating that using a program.

By the way, $t_\infty \sim N(0,\ 1)$.