STA238 Notes

Statistics for real

https://github.com/ICPRplshelp/

Last updated January 19, 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 can view random samples as **unbiased samples.** In a perfect world, statistical methods like these work perfectly, but garbage in, garbage out.
- 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 using case 3 must be given such that doing this is possible.

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

The t-distribution is symmetric around 0. To read the t-distribution table, focus on the degrees of freedom that you want to target. If we happen to know the area of one side of the tail, we can find the critical points $t_{\text{corresponding tail area}}$. Look for the **one-tailed probability** equal to the area under the tail of the curve. That is the critical value; value on the x-axis.

Hence, $t_{0.025,\ 11} = 2.201$.

The t-curve is symmetric, so the area below the left and the right tails should match. When we add the area below both tails, we get the area of one tail multiplied by two.

1.4 Chi-Squared Distribution

$$\mu \to \bar{y} \to N \text{ or } t$$
$$p \to \widehat{p} \to N$$
$$\sigma^2 \to s^2 \to \chi^2$$

We want to conduct statistical inferences for μ. $\bar{y}$ is the candidate. We have two options:

- Normal distribution
- t-distribution

Which depends on the sample size. If we are interested on the population proportion, we use sample proportion $\widehat{p}$, which follows the normal distribution **under certain conditions (if it fails, we can't answer anything, not even use the t-distribution).**

The candidate for population variance σ^2 is s^2.

$$y_1,\ y_2,\ \ldots,\ y_n \sim N(\mu,\ \sigma)$$
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Which is the definition of the sample variance. We can use this to make statistical inferences of σ^2. We can't find σ^2 directly, and we **cannot** find the distribution for s^2 directly, but we can find the distribution for:

$$\frac{(n-1)s^2}{\sigma^2} \quad \text{this quantity}$$

And it follows this distribution:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1=df}$$

We can prove this result using the moment generating function, but we're not going to use it here.

For this distribution, what we are sampling from **must be normally distributed.**

We'll use this result to conduct statistical inferences for σ^2.

1.4.1 Examples of Chi-2

You can't pour exactly 500mL of water into a bottle each time and get it perfectly. So, each of your water buckets, which you aimed to pour water in, will hold water. The quantity of water held for each bucket will be a random variable Y.

Better to use a confidence interval, right?

You might get a question looking like this, where you want to find b_1 and b_2:

$$P\left(b_1 \leq s^2 \leq b_2\right) = 0.9$$

To find b_1 and b_2, we must know the distribution of s^2. But we don't know, so we'll set up for that quantity. Multiply both sides by $(n-1)$ and divide it by σ^2, so the probability does not change.

$$P\left(\frac{(n-1)b_1}{\sigma^2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \frac{(n-1)b_2}{\sigma^2}\right) = 0.9$$

If we let $n = 10$ and $\sigma^2 = 1$:

$$P\left(9b_1 \leq \chi^2_9 \leq 9b_2\right) = 0.9$$

The χ^2 distribution is skewed to the right, so it is not symmetric. We have $df = 9$.

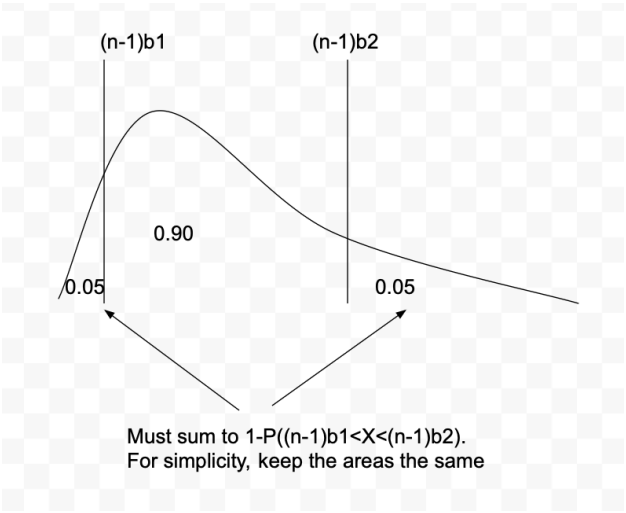

Figure 1: Chi-2 density visualization

If we focus on $df = 9$, at the **row** of the chi-squared table:

$$(n-1)b_2 = 16.919 = \chi^2_{0.05}$$
$$(n-1)b_1 = 3.325 = \chi^2_{0.95}$$

From this, we can conclude that (recall $n = 10$):

$$b_1 = 0.369$$
$$b_2 = 1.88$$

And we've built a 90% confidence interval on s^2. If we don't know what σ^2 is, then our confidence interval will likely contain σ^2 as a term.

1.4.2 (Optional) Why the above works

A bit of a proof for the above

$$
\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
\frac{(n-1)s^2}{\sigma^2} &= \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sigma^2} \\
&= \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2} - \frac{(\bar{y} - \mu)^2}{\frac{\sigma^2}{n}}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{y_i - \mu}{\sigma} \sim N(0,\ 1) \\
&\left(\frac{y_i - \mu}{\sigma}\right)^2 \sim \chi^2_1 \quad \text{squaring both sides, not proving this}
\end{aligned}
$$

You won't be tested on this, though.

1.5 F-Distribution (Statistical Inferences for Two Populations)

If we have a bunch of students and we have $Y \rightarrow$ height for each student. We can't say that all students are the exact same. We'll have to separate gender apart (one of the covariate information), age, and so on.

If we split students into

- Population 1: F student population
- Population 2: M student population

This means we can divide students into groups. In statistics, all assumptions are ideally wrong, but when we use some assumptions, we can make some real conclusions for usage.

If we assume that all F and M students have the same characteristics (i.e., all have the same age), and

- Population 1 has mean $\mu_1,\ \sigma_1^2$
- Population 2 has mean $\mu_2,\ \sigma_2^2$ (population mean and variance respectively, for M and F student height)

What if we want to calculate statistical inference for $\mu_1 - \mu_2$, or statistical inference for $\frac{\sigma_1^2}{\sigma_2^2}$?
In hypothesis testing:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad H_A : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

If I want to do statistical inference for $\frac{\sigma_1^2}{\sigma_2^2}$, this is an unknown ratio. But the corresponding statistics is

$$\frac{s_1^2}{s_2^2}$$

It means that we can collect samples from the first population and the second population. So:

- For the F group, we have $n_1,\ \bar{y}_1,\ s_1^2$
- For the M group, we have $n_2,\ \bar{y}_2,\ s_2^2$

If we know the distribution of $\frac{s_1^2}{s_2^2}$, then we can make statistical inferences with $\frac{\sigma_1^2}{\sigma_2^2}$.
We'll need to start with the chi-squared distribution first. Firstly, we assume that:

$$W_1 \sim \chi^2_{\nu_1 \leftarrow df,\ \text{not necessairly int}}$$
$$W_2 \sim \chi^2_{\nu_2}$$

Both are random variables, so they take some values in a particular range. More importantly, W_1 and W_2 are **independent**. Under these assumptions, we can do statistical inference. If that is the case (note that ν_1 is the degrees of freedom):

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F^{\nu_1 \leftarrow \text{numerator}}_{\nu_2 \leftarrow \text{denominator } df} \text{ or } F_{\nu_1,\, \nu_2}$$

Here, F is said to have an F distribtuon with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom.

The reason why W is divided by ν because W already is multiplied by ν and we need to cancel out its effects.

This is related to ANOVA. ANOVA (analysis of variance) has applications in agriculture when we consider which fertilizer is good, which ones aren't.

1.5.1 Examples

The F distribution looks like it is skewed to the right.

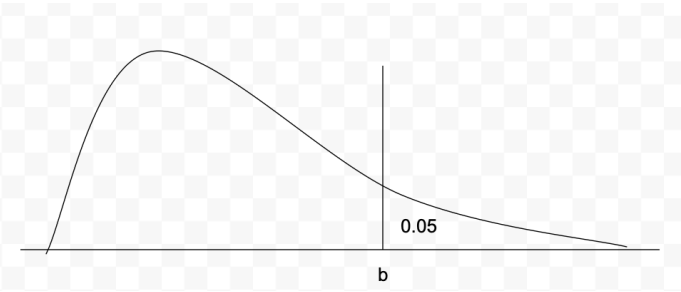

Figure 2: F-distributed density function

Your questions will probably look like

$$P(F \leq b) = 0.95$$

Good luck using the table here. Consult the STA237 notes on how to use the table. Again, you'll be comparing sample variances.

2 Point Estimation

Expect a question on this

ESTIMATORS

Expectation means population mean. So:

$$E(\bar{y}) = \mu$$

What I'm going to do: estimate the mean.

$$\underset{\text{estimator of } \mu}{\widehat{\mu}} = \bar{y}$$

For example, what is the average height of the Canadian population? From a sample, we might estimate that it is 5.5ft from a 1000-person random sample.

μ is still unknown, and it won't be equal to exactly 5.5ft.

$\bar{y}$ is the unbiased estimate of μ.

When we define $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$, we can prove that

$$E\left(s^2\right) = \sigma^2 \quad \text{what did you expect?}$$

So, the sample variance is an unbiased estimator for σ. So:

$$\underset{\substack{\text{estimator for} \\ \text{variance}}}{\widehat{\sigma}^2} = s^2$$

BIASED ESTIMATOR (the evil version of the unbiased estimator)

$$s_*^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

THE DIFFERENCE IS THAT IT ISN'T AN UNBIASED ESTIMATOR. If you define sample variance using this formula. So, s_*^2 is a **biased** estimator for σ^2.

2.1 The General Parameter θ

Now, we've learned about the following statistics:

- $\bar{y} \leftrightarrow \mu$
- $\widehat{p} \leftrightarrow p$, the population proportions
- $s^2 \leftrightarrow \sigma^2$

The general parameter θ MIGHT be any of $\mu,\ p,\ \sigma^2$ (think of interfaces, abstraction, polymorphism… could be any, and I don't know which one yet).

$\widehat{\theta}$ is the point estimator for the parameter θ.

$\widehat{\theta}$ is an unbiased estimator if $E\left(\widehat{\theta}\right) = \theta$. Otherwise, it is said to be a biased estimator.

Now go ahead and replace θ and $\widehat{\theta}$ with $(\mu, \bar{y}),\ \left(s^2, \sigma^2\right),\ (\widehat{p}, p)$ respectively.

2.2 Bias of the Point Estimator

The bias of a point estimator $\widehat{\theta}$ is:

$$B\left(\widehat{\theta}\right) = E\left(\widehat{\theta}\right) - \theta$$

When considering $\bar{y}$ and μ, what is the bias?

$$B\left(\bar{y}\right) = E\left(\bar{y}\right) - \mu = \mu - \mu = 0$$

And there, we can say that $\widehat{y}$ is an **unbiased estimator.**

2.3 Mean Squared Error

$$\begin{aligned} MSE\left(\widehat{\theta}\right) &= E\left(\left(\widehat{\theta}-\theta\right)^2\right) \\ &= V\left(\widehat{\theta}\right)+\left(B\left(\widehat{\theta}\right)\right)^2 \end{aligned}$$

No need to prove this. However, the variance for $\widehat{\theta}$ is…

Well, look at $V(\bar{y}) = E\left((\bar{y}-E(\bar{y}))^2\right)$. So:

$$E\left(\underbrace{\left(\widehat{\theta}-E\left(\widehat{\theta}\right)\right)}_{\text{variance}}+\underbrace{\left(E\left(\widehat{\theta}\right)-\theta\right)}_{\text{bias}}\right)^2$$

If we expand this, we get $MSE\left(\widehat{\theta}\right) = V\left(\widehat{\theta}\right)+B\left(\widehat{\theta}\right)^2$. Don't prove it.

3 Confidence Intervals

We can estimate $\widehat{\mu}$ using $\bar{y} = \underbrace{5.5}_{\text{example}}$. We want an interval estimate such as a 95% confidence interval for $\mu = \underbrace{(4.8,\ 5.9)}_{\text{example}}$. So, there is a 5% possibility that we are out of the interval (such as a 5% error).

An interval estimator is a rule specifying for using the sample measurements to calculate two number that form the endpoints of an interval.

The interval should capture the true population parameter (be within), such as the population mean.

An interval that is too wide gives us no information. A 100% confidence interval doesn't give us any meaning, so we'll consider 95%, 99%, or 90% confidence intervals. We can allow some error in the CI, and the **error is called the level of significance** α.

$$\alpha = 0.05$$

The confidence level is $1-\alpha$. The confidence the unknown population parameter will be in that interval.

- Lower bound is lower confidence limit
- Upper bound is upper confidence limit

3.1 Large-Sample CIs

Take $\bar{y}$ and μ. Consider a large sample, the sample size ≥ 30. By CLT, we can say that $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. We can standardize: $Z = \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,\ 1)$.

We want to construct a $1-\alpha$ confidence interval, and we'll going to make a two-sided confidence interval, so we're going to distribute error equally on the tails of the normal curve.

BUILDING THE CONFIDENCE INTERVAL

Then, we get critical values which we would put on the x-axis:

- RIGHT: $Z_{\frac{\alpha}{2}}$
- LEFT: $-Z_{\frac{\alpha}{2}}$

Because we have $-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}$

We want the random variable in the middle to be set to

$$-Z_{\frac{\alpha}{2}} < \frac{\bar{y}-\mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\frac{\alpha}{2}}$$

Multiply both sides by $\frac{\sigma}{\sqrt{n}}$

$$-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{y}-\mu < Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

And move $\bar{y}$ to adjacent sides:

$$-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}-\bar{y}<-\mu<Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}-\bar{y}$$

Multiply all sides by -1:

$$\bar{y}-\frac{Z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}<\mu<\bar{y}+\frac{Z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}$$

And there, we've got our $100(1-\alpha)\%$ confidence interval for $\mu=\bar{y}\pm\frac{Z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}$ (that is the uncertainty). Which is:

$$\bar{y}\pm\left(Z_{\frac{\alpha}{2}}SE\left(\bar{y}\right)\right)$$

This is the general result. If we want to do this for population proportion, just take the mean and variance and plug in stuff from there.

If you need to use a t-distribution yet, replace Z with t. If σ is unknown, but we have a large sample, replace σ with s.

3.1.1 Confidence Intervals for Population Proportion

To form a $100(1-\alpha)\%$ CI for an unknown proportion p ($\widehat{p}$ can either be normal or unknown):

- Is $\widehat{p}\pm Z_{\frac{\alpha}{2}}SE\left(\widehat{p}\right)$, which is the CI
- $SE\left(\widehat{p}\right)=\sqrt{\frac{p(1-p)}{n}}$

We can't make a confidence interval if we need an unknown parameter. We're going to estimate it:

$$\widehat{SE}\left(\widehat{p}\right)=\widehat{\sigma}_{\widehat{p}}=\sqrt{\frac{\widehat{p}\left(1-\widehat{p}\right)}{n}}$$