Statistics

MATH 324

McGill University, Montréal, Canada

Fall 2018

Sampling distributions

- Recall: Statistics is the science of extracting information from data using tools from mathematics, in particular, probability.

1- In this chapter, we formally define a statistic.

2- Introduce the distribution of a statistic: sampling distribution.

3- The Central Limit Theorem (CLT), and some related topics.

Statistic

- Let $\underline{X} = (X_1, ..., X_n)$ be a random sample from some distribution F.

  Definition:

  A **statistic** is a function of only the random sample and some known constants:

  $$T(\underline{X}) = T(X_1, ..., X_n) : \mathcal{X} \longrightarrow \mathbb{R}^d.$$

  where $\mathcal{X} \subset \mathbb{R}^n$ is referred to as the sample space, and $d \geq 1$.

- Statistical analyses use various **statistics** for various purposes.

- Note:

  One assumption we often (but not always) make is that the random variables $X_1, X_2, \ldots, X_n$ are a random sample, i.e. that they are independent and identically distributed according to the same probability distribution, say, F.

Examples

- Sample mean (average):

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Sample variance:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

- Order statistics:

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}.$$

- Range:

$$R_n(\underline{X}) = X_{(n)} - X_{(1)}.$$

Remarks

- A **statistic** is itself a random variable; hence, it has a distribution.

  Defintion:

  The distribution of a **statistic** is called sampling distribution.

- Example 7.1: an illustrative example on sampling distribution.

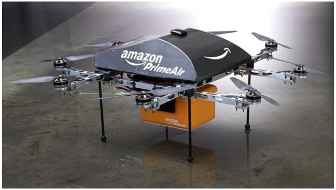
More on sampling distribution

- It depends on the underlying distribution F from which the random sample $X_1, \ldots, X_n$ is taken.
- It depends on the statistic $T(\underline{X})$ under consideration.
- It depends on the sample size n.
- It may or may not be computed explicitly.

Why do we even care about the sampling distribution?

- *CBC News Post: Mar 30, 2015*:

  Seattle-based Amazon wants to deliver packages of under five pounds in 30 minutes or less using its Amazon Prime Air autonomous drones in the near future.



- Consider the timing for $n = 100$ deliveries, with observed average $\bar{x}_n = 33$ minutes. What can we conclude from this observation? Is the mean delivery time higher than what is claimed?

- Let X be the delivery time of a randomly selected Amazon Prime Air autonomous drone (the type for which data is collected). The distribution of X is denoted by F.

- A random sample: $X_1, X_2, \ldots, X_n$ are iid from F.

- The sample average: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

- Based on the post-experimental data, we have observed that $\overline{x}_n = 33$ minutes, with $n = 100$. We would like to see, if the company's claim is true, how likely is to observe such sample average !

- To answer this question, by using statistics and probability language, we need certain tools that we discuss now.

Samples of Gaussian random variables

Assumption: in this Sub-section, we assume that $X_1, X_2, \dots, X_n$ is a random sample from $N(\mu, \sigma^2)$; (unless otherwise is stated).

- The Normal distribution fits reasonably well to many data sets, and is a suitable approximation to many discrete and continuous distributions.
- Compared to other distributions, it is easier to work with the normal distribution in many statistical analysis problems.

A. Sampling distribution of the sample average

Recall the following result:

- Theorem 6.3:

  Let $X_1, \ldots, X_n$ be independent random variables, where $X_i \sim N(\mu_i, \sigma_i^2)$, $\mu_i = E(X_i), \sigma_i^2 = Var(X_i)$, for $i = 1, \ldots, n$. Then,

$$Y_n = \sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right),$$

  $a_1, \ldots, a_n \in \mathbb{R}$ are constants.

Important special case

- Set $a_1 = \ldots = a_n = \frac{1}{n}$, and
$$\mu_1 = \ldots = \mu_n = \mu \quad , \quad \sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2.$$
Then, $Y_n = \overline{X}_n$. Furthermore,

- Theorem 7.1:

  Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$, then
$$\overline{X}_n \sim N(\mu, \sigma^2/n).$$

- PROOF. Use Theorem 6.3.

Remarks

- For any sample size n, we have

$$E(\overline{X}_n) = \mu \quad , \quad Var(\overline{X}_n) = \frac{\sigma^2}{n}.$$

- For any sample size n,

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

- As $n \to \infty$: $Var(\overline{X}_n) \to 0$.

  This implies that $\overline{X}_n$ converges in probability to μ as the sample size n increases.

Drone example (cont'd...)

- Assuming that the delivery time X has a normal distribution $N(\mu, \sigma^2)$, we have that:

$$\overline{X}_{100} \sim N(\mu, \frac{\sigma^2}{100}).$$

- Note: the above sampling distribution depends on two unknown parameters:

$$(\mu, \sigma^2).$$

- For now, we cannot proceed unless we make more assumption(s)!

Some applications of the sampling distribution of $\overline{X}_n$

To see why it is useful to know a sampling distribution, make the following assumption (for the time being, out of convenience only):

- Assumption:

The true value of the standard deviation is $\sigma = 5$ minutes. Thus,

$$\overline{X}_{100} \sim N(\mu, \frac{25}{100}).$$

1. Proving or disproving a claim about μ

- Suppose Amazon's claim is true and $\mu = 30$. What is the probability of observing a random sample with average delivery time at least 33 minutes?

$$\Pr(\overline{X}_{100} \geq 33) = \Pr\left(Z \geq \frac{33-30}{5/10}\right) = \Pr(Z \geq 6) \approx 9.9 \times 10^{-10}$$

- Based on this data, the claim is thus very unlikely to be true!

  This type of argument is called argumentum ad absurdum.

2. Finding a plausible range of values for μ

- What is the chances that $\overline{X}_{100}$ lies within 1 minute from the true average delivery time (μ)?

$$\Pr(|\overline{X}_{100} - \mu| \leq 1) = ??? \approx 0.9545$$

3. Determining a minimum sample size

- Suppose we want to be 90% sure that $\overline{X}_n$ is within 1 minute from μ (90% is 18 out of 20). How many timing (n) we should test?

$$P(|\overline{X}_n - \mu| \leq 1.0) = 0.90.$$

- Note that if $Z \sim N(0, 1)$, then $P(|Z| \leq 1.645) \approx 0.90$. Hence,

$$\frac{1.0}{5/\sqrt{n}} \geq 1.645 \Longleftrightarrow n \geq \frac{5^2 \times (1.645)^2}{1.0^2} = 68.0625.$$

B. Sampling distribution of the sample variance

Recall the following result:

- Theorem 6.4:

  Let $X_1, \ldots, X_n$ be independent random variables, where $X_i \sim N(\mu_i, \sigma_i^2)$, $\mu_i = E(X_i), \sigma_i^2 = Var(X_i)$, for $i = 1, \ldots, n$. Define,

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}.$$

  Then, $Z_1, \ldots, Z_n$ are independent and they all have the same distribution $N(0, 1)$. Also, $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}$.

- Special case: $\mu_1 = \ldots = \mu_n = \mu$ and $\sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2$

Sampling distribution of S_n^2 (cont'd...)

- Theorem 7.3:

  Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Then,

  $$\frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \sim \chi^2_{(n-1)}.$$

  Moreover, $\overline{X}_n$ and S_n^2 are independent.

- PROOF. Will be presented later on when we have enough tools.

- Compare Special case of Theorem 6.4 and Theorem 7.3. Pay attention to the degrees of freedoms of the two χ^2 distributions.

Some properties of S_n^2

- Note that: $E\{\chi^2_{(r)}\} = r \;\; , \;\; Var\{\chi^2_{(r)}\} = 2r.$

- Then, it is easy to see that:

$$E\{S_n^2\} = \sigma^2 \;\; , \;\; Var(S_n^2) = \frac{2\sigma^4}{n-1}$$

- Hence, $S_n^2 \xrightarrow{p} \sigma^2$, as $n \to \infty$.

- This implies that as the sample size grows larger, the sample variance S_n^2 will be closer and closer to the population variance σ^2.

Drone example: cont'd...

- What is the probability that the ratio $\frac{S_n^2}{\sigma^2}$ lies in $[0.7, 1.3]$?

$$\Pr\left\{0.7 \leq \frac{S_n^2}{\sigma^2} \leq 1.3\right\} = \Pr\left\{69.3 \leq \frac{(n-1)S_n^2}{\sigma^2} \leq 128.7\right\} = 0.9658.$$

- The above calculation implies that we are "96.58%" confident that σ^2 belongs to the interval $[S_n^2/1.3, S_n^2/0.7]$.

- For example, assume that the observed value of the sample standard deviation is $s_n = 4.5$. Then,

$$[s_n^2/1.3, s_n^2/0.7] = [15.58, 28.93].$$

Be careful about the interpretation of this interval.

The student distribution

- Definition 7.2:

  Suppose $Z \sim N(0,1)$ and $W \sim \chi^2_{(\nu)}$ are independent. Then,

  $$T = \frac{Z}{W/\sqrt{\nu}} \sim t_{(\nu)}.$$

  we say T has a Student t distribution with ν degrees of freedom.

- Its pdf has a complex form and we do not directly use it in this course.

- This distribution is due to William S. Gosset, who published it under the pen name "Student" (he worked for Arthur Guinness & Son, Dublin).

Construction of the Student t distribution

- Theorem:

  Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Then,

  $$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \sim t(n-1).$$

- PROOF. Use Theorems 7.1 and 7.3, and Definition 7.2.

1. Drone example (cont'd...)

- We revisit the calculations on page 15 of the notes. Here, we do not know σ^2. Then,

$$\Pr(\overline{X}_{100} \geq 33) = \Pr\left(T \geq \frac{33 - 30}{4.5/10}\right) = \Pr(T \geq 6.67) = 7.4 \times 10^{-10}$$

  where $T \sim t_{(99)}$.

- Again, based on this data, their claim seems very unlikely to be true!

McGill

2. Finding a plausible range of values for μ

- What is the chances that $\overline{X}_{100}$ lies within 1 minute from the true average delivery time (μ)? (we do not know σ^2).

$$\Pr(|\overline{X}_{100} - \mu| \leq 1) = ??? \approx 0.9713$$

C. Sampling distribution of the ratio of two sample variances

- Let $X_1, \ldots, X_n \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2)$ be two independent random samples.
- Question: how do we compare the two variances σ_1^2 and σ_2^2?

The F statistic

- Definition: Let $W_1 \sim \chi^2_{(\nu_1)}$ and $W_2 \sim \chi^2_{(\nu_2)}$ be independent random variables. Then,

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

is said to have the Fisher-Snedecor F distribution with (ν_1, ν_2) degrees of freedom, and we write $F \sim F_{(\nu_1,\nu_2)}$.

- Note:

$$F \sim F_{(\nu_1,\nu_2)} \Longleftrightarrow \frac{1}{F} \sim F_{(\nu_2,\nu_1)}$$

- Similar to other well known distributions, the quantiles of the F distribution can be obtained from statistical tables. The F distribution is also available in R.

Comparing sample variances

- Theorem

  Consider the independent random samples $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_1, \ldots, Y_m \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$. Then

  $$\frac{S_n^2/\sigma_1^2}{S_m^2/\sigma_2^2} \sim F_{(n-1, m-1)}.$$

- PROOF. To be discussed in class.

What if normality assumption does not hold?

- In our discussion in the last two lectures, we have been assuming that the data generating mechanism is a Gaussian distribution.
- The assumption led to convenient well-known distributions for the sample mean, variance, etc.
- Let us relax the normality assumption and see what we can do.

The Central Limit Theorem (CLT)

- Theorem 7.4: Let $X_1, \ldots, X_n$ be a random sample from an arbitrary distribution F with $E(X_i) = \mu$ and $0 < Var(X_i) = \sigma^2 < \infty$. Define

$$U_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}.$$

Then as $n \to \infty$, for all $x \in \mathbb{R}$,

$$G_n(x) = \Pr(U_n \leq x) \longrightarrow \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

- We say U_n converges in distribution to $N(0, 1)$, and we write $U_n \xrightarrow{d} N(0, 1)$.

Reality check

- Under the assumptions of Theorem 7.4,

$$
\begin{aligned}
E(\overline{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu \\
Var(\overline{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}.
\end{aligned}
$$

- This means, the U_n in Theorem 7.4 is designed so that

$$E(U_n) = 0 \ , \ \ Var(U_n) = 1.$$

Remarks

- Practical implication of Theorem 7.4: for large sample sizes n,

$$U_n \approx N(0,1) \text{ , or equivalently } \overline{X}_n \approx N(\mu, \frac{\sigma^2}{n}),$$

  where "$\approx$" means "approximation".

  This is irrespective of the underlying distribution F, as long as $0 < Var(X_i) = \sigma^2 < \infty$.
- The approximation becomes arbitrarily good, as n grows. The speed at which this occurs depends on F, though.
- It has been generalized in various ways, e.g., by Lindeberg and Lévy, Lyapunov, etc.

Drone example: revisited

- Recall the calculations on pages 15, 16, and 17. Under the normality assumption of the distribution of delivery time, the probability calculations were exact.

- Now, let us relax the assumption that the delivery time, as a random variable, follows a normal distribution $N(\mu, \sigma^2)$. That means, it has an unknown distribution F.

- Repeat all the calculations, except that the probability statements will all be approximations.

- Note: the sample size $n = 100$ is large enough, and hence the approximation based on the CLT is very good.

Next:

- Can we also relax the assumption of "known variance σ^2" in Drone example calculations?

- Answer: YES.

- We will use Slutsky's Theorem:

  Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be two sequences of random variables such that as $n \to \infty$, $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, for some constant c. Then, as $n \to \infty$

  1. $X_n + Y_n \xrightarrow{d} X + c$
  2. $X_n \times Y_n \xrightarrow{d} X \times c$
  3. If $c \neq 0$, $X_n / Y_n \xrightarrow{d} X / c$.

- Theorem:

  Let $X_1, X_{,} \ldots, X_n$ be a random sample from an arbitrary distribution F such that $E(X_i^4) < \infty$. Then, as $n \to \infty$,

$$W_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \xrightarrow{d} N(0, 1).$$

- PROOF. Will be discussed in class.

- NOTE:

  If $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then W_n has an exact t-student distribution with $(n-1)$ degrees of freedom. If $n \to \infty$, then $t_{(n-1)} \xrightarrow{d} N(0, 1)$.

Drone example: revisited

- Recall the calculation on page 24.

$$\Pr(\overline{X}_{100} \geq 33) = \Pr\left(U_{100} \geq \frac{33-30}{4.5/10}\right) = \Pr(U_{100} \geq 6.67)$$

$$\approx 1 - \Phi(6.67) = 1.28 \times 10^{-11}.$$

- Again, their claim seems very unlikely to be true!

- What is the chances that $\overline{X}_{100}$ lies within 1 minute from the true average delivery time (μ)?

$$\Pr(|\overline{X}_{100} - \mu| \leq 1) = ??? \approx 0.9736.$$

McGill

The Normal approximation to the binomial distribution: (CLT)

- Will be discussed in class.