LECTURE 5: PROPERTIES OF A RANDOM SAMPLE

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG
SEPTEMBER 22, 2021

1. Random Samples

Suppose n number of observations are randomly sampled from the density $f(x)$. The first observation is X_1, the second observation X_2, and so on. The outcome is a *random sample* described by the random variables $X_1, \ldots, X_n$.

Random sampling is the default assumption that we maintained throughout.[1] It means that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d) with the pdf $f(x)$. That is, the density of each $X_1, \ldots, X_n$ is the same and equals $f(x)$. Moreover $X_1, \ldots, X_n$ are mutually independent.

The joint pdf of $X_1, \ldots, X_n$ is:

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n) = \Pi_{i=1}^{n} f(x_i)$$

1.1. Statistic

Let $X_1, \ldots, X_n$ be a random sample from a population. Let $T(X_1, \ldots, X_n)$ be a function that maps the sample space of $(X_1, \ldots, X_n)$ into $\mathbb{R}$. Then $Y = T(X_1, \ldots, X_n)$ is called a statistic. A statistic is a random variable. The pdf of $Y = T(X_1, \ldots, X_n)$ is called the *sampling distribution* of Y.

For example, the *sample mean* is the statistic defined by:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

The *sample variance* is the statistic defined by:

[1] If our survey and experiment is not randomized correctly, then random sampling is violated. For instance, when a certain type of individuals opt out of the survey. The population also needs to be large enough, so that surveying the first household (and hence "removing" this household from the population to be surveyed next) has no effect on the population.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

1.2. Unbiased estimator

A statistic $T(X_1, \ldots, X_n)$ is an unbiased estimator of a population parameter θ if $\mathbb{E}[T(X_1, \ldots, X_n)] = \theta$.

Let $\mathbb{E}[X_i] = \mu$, then $\bar{X}$ is an unbiased estimator (or statistic) of μ.

$$\begin{aligned}
\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n}\,\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] \\
&= \frac{1}{n} n\mu = \mu
\end{aligned}$$

Moreover, the statistic $\bar{X}$ has a variance of $\frac{\sigma^2}{n}$, where $\mathrm{Var}(X_i) = \sigma^2$.

$$\begin{aligned}
\mathrm{Var}[\bar{X}] &= \mathrm{Var}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \\
&= \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

The sample variance S^2 is an unbiased estimator of σ^2.

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\bar{X}^2\right] \\
&= \frac{1}{n-1}\,\mathbb{E}\left[\sum_{i=1}^{n}X_i^2 - 2n\bar{X}^2 + n\bar{X}^2\right] \quad \text{using } \sum X_i = n\bar{X} \\
&= \frac{1}{n-1}\left(n\,\mathbb{E}[X_i^2] - n\,\mathbb{E}[\bar{X}^2]\right)
\end{aligned}$$

Since we have $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2$, and $\mathbb{E}[\bar{X}^2] = \mathrm{Var}(\bar{X}) + \mathbb{E}[\bar{X}]^2 = \sigma^2/n + \mu^2$,

$$\begin{aligned}
\mathbb{E}[S^2] &= \frac{n}{n-1}(\sigma^2+\mu^2) - \frac{n}{n-1}\left(\frac{\sigma^2}{n}+\mu^2\right) \\
&= \frac{n}{n-1}\sigma^2 - \frac{1}{n-1}\sigma^2 \\
&= \sigma^2
\end{aligned}$$

The bias of a statistics with respect to the population parameter θ is $\mathbb{E}[T(X_1,\ldots,X_n)] - \theta$. For instance, if we were to use the sample variance formula $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$, then:

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

Therefore the bias in using the statistic $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ in estimating the population parameter σ^2 is $\frac{1}{n}\sigma^2$, which diminishes as the sample size, n increases.

2. Sampling distribution

Let $X_1,\ldots,X_n$ be a random sample from a $N(\mu,\sigma^2)$ distribution. Consider the sample mean $\bar{X} = \frac{X_1+\cdots+X_n}{n}$, which is a random variable. The sampling distribution of $\bar{X}$ is Normally distributed as $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. In another words, the pdf of $\bar{X}$ is $\mathcal{N}(\mu, \frac{\sigma^2}{n})$. This can be proven using moment generating functions.

Sampling distribution is important because it describes the distribution of an estimator (or statistic) due to sampling variation. For example, the sample mean we calculated from a sample is a noisy estimate of the true population mean, through the sampling distribution, we get a sense of how noisy this estimate is.

Now let S^2 be the sample variance (as defined above). The sampling distribution of $(n-1)S^2/\sigma^2$ is a chi squared distribution with $n-1$ degrees of freedom.

We now use simulations to derive the sampling distributions and check that the above statements are correct.

2.1. Chi squared distribution

Notation: χ^2_p denotes a chi squared distribution with p degrees of freedom.

If X is a $N(0,1)$ random variable, then $X^2 \sim \chi^2_1$. If $X_1, \ldots, X_n$ are independent $N(0,1)$, then $Z = X_1^2 + \cdots + X_n^2$ is distributed according to χ^2_n. The pdf of χ^2_p is $f(x) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{\frac{p}{2}-1} e^{-\frac{x}{2}}$, for $x > 0$. Recall that $\Gamma(\cdot)$ is the Gamma function. In fact, χ^2_p is a special case of the Gamma distribution.

The mean of χ^2_p is p, while the variance of χ^2_p is $2p$. Visualize the chi squared distribution using Mathematica's:

```
PDF[ChiSquareDistribution[p],x]
```

```
Manipulate[Plot[PDF[ChiSquareDistribution[p],x],{x,0,20}],{p,1,5}].
```

The sample variance is $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$. Since $(n-1)S^2/\sigma^2$ is a chi-squared distribution with $n-1$ degrees of freedom, we have:

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

As such $\mathbb{E}[S^2] = \sigma^2$ and $\text{Var}(S^2) = 2\sigma^4/(n-1)$.

To see intuitively why the sample variance has a χ^2_{n-1} distribution:

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

:

If we know that the population mean is μ, then we have $\frac{X_i-\mu}{\sigma} \sim N(0,1)$, and as such $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$. Since we do not know μ and estimate μ by imposing that $\mu = \frac{X_1+\cdots+X_n}{n}$, we lose one degree of freedom.

2.2. Student's t distribution

Recall that the sampling distribution of $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$. Therefore, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

What if we replace σ with the sample standard deviation S? What is the sampling distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$, where S is the sample standard deviation?

Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. The statistic $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has *Student's t distribution with $n-1$ degrees of freedom.*[2]

Let us sketch the derivation of the pdf of the Student's t distribution. By multiplying both the numerator and denominator by σ, we get

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} \tag{1}$$

The numerator $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ is distributed $N(0,1)$. While the denominator $\sqrt{S^2/\sigma^2}$ is distributed as $\sqrt{V/(n-1)}$ where $V \sim \chi^2_{n-1}$. Moreover, the sample variance and mean are independent (see Theorem 5.3.1 in Casella-Berger). Therefore, the Student's t distribution with $n-1$ degrees of freedom is distributed as $\frac{U}{\sqrt{V/n-1}}$, where $U \sim N(0,1)$ and $V \sim \chi^2_{n-1}$, with U and V independent. The pdf of the Student's t distribution can then be obtained by applying the bivariate change-of-variables formula.

The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean. As the degree of freedom $n-1$ increases, the t-distribution converges in distribution to the standard Normal.

The pdf of the t-distribution with degree of freedom v is:

$$f_v(x) = \frac{\left(\frac{v}{v+x^2}\right)^{\frac{v+1}{2}}}{\sqrt{v}Beta\left(\frac{v}{2}, \frac{1}{2}\right)}$$

It has mean (median, model) of zero, but its variance is $\frac{v}{v-2}$.

[2] If σ^2 were known, then the sampling distribution of $\bar{X}$ would be known at a given hypothesized value of μ. In practice, σ^2 is not known. William Gosset (under the pseudonym of Student) in the 1900s proposed substituting the unknown σ^2 with the sample variance S^2, and derived the resulting sampling distribution.

3. Order Statistics

Let $X_1, \ldots, X_n$ be a random sample from a population with distribution $f(x)$.

Let $X_{(n)}$ be the largest value from the sample $X_1, \ldots, X_n$. That is, $X_{(n)} = \max_{1 \le i \le n} X_i = \max(X_1, \ldots, X_n)$. The random variable $X_{(n)}$ is called the largest order statistics. While $X_{(1)} = \min(X_1, \ldots, X_n)$ is called the first-orders statistics.

What is the distributions of $X_{(1)}$ and $X_{(n)}$? Order statistics plays a big role in Auction Theory. In a second-price auction, the participant with the largest bid wins the auction but pays the second-largest bid. If bids are assumed to be realized i.i.d from a distribution, then the revenue from the auction is distributed as the second-largest order statistics.

To obtain the distributions of the order statistics, we start with the cdf.

$$\begin{aligned} P(X_{(n)} \le x) &= P(\max(X_1, \ldots, X_n) \le x) \\ &= \text{Probability that all } n \text{ observations are smaller than } x \\ &= F(x)^n \end{aligned}$$

Hence the cdf of $X_{(n)}$ is $F(x)^n$, and the pdf is $nF(x)^{n-1}f(x)$.

The smallest-order statistic is:

$$\begin{aligned} P(X_{(1)} \le x) &= P(\min(X_1, \ldots, X_n) \le x) \\ &= 1 - P(\min(X_1, \ldots, X_n) > x) \\ &= 1 - \text{Probability that all } n \text{ observations are larger than } x \\ &= 1 - (1 - F(x))^n \end{aligned}$$

Which leads to the density $n(1 - F(x))^{n-1}f(x)$.

In general, the density of $X_{(k)}$ can be derived as follows:

$$\begin{aligned} P(X_{(n-1)} \le x) &= \text{Probability that exactly } n-1 \text{ observations are smaller than } x \\ &\quad + \text{Probability that exactly } n \text{ observations are smaller than } x \\ &= \binom{n}{n-1} F(x)^{n-1}(1 - F(x)) + F(x)^n \end{aligned}$$

$$
\begin{aligned}
P(X_{(n-2)} \leq x) = & \text{ Probability that exactly } n-2 \text{ observations are smaller than } x \\
& + \text{Probability that exactly } n-1 \text{ observations are smaller than } x \\
& + \text{Probability that exactly } n \text{ observations are smaller than } x \\
= & \binom{n}{n-2} F(x)^{n-2}(1-F(x))^2 + \binom{n}{n-1} F(x)^{n-1}(1-F(x)) + F(x)^n
\end{aligned}
$$

3.1. Order statistic for the Uniform distribution

Suppose $X_1, \ldots, X_n$ is a random sample from the Uniform distribution $U[0,1]$.

The largest-order statistic has the pdf $f(x) = nx^{n-1}$ with the support on $x \in (0,1)$. Further, we can calculate the expectation with respect to this density, which is $\mathbb{E}[X_{(n)}] = \frac{n}{n+1}$. Therefore, when $n = 20$, the maximum out of n uniform random values is $20/21 \approx 0.952$, *on average*. In fact, $f(x) = nx^{n-1}$ is the Beta distribution with parameters $\alpha = n$ and $\beta = 1$. In general, the k-th order statistic of a Uniform distribution is $Beta(k, n+1-k)$.

As an exercise, show (using simulations) that the largest-order statistic of the Standard Normal also has an expectation that is concave in terms of n, the sample size.

4. Convergence and consistency

4.1. Convergence in probability

A sequence of random variables $Y_1, Y_2, \ldots$, converges in probability to a number θ if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|Y_n - \theta| \geq \epsilon) = 0$$

or equivalently, $\lim_{n\to\infty} P(|Y_n - \theta| < \epsilon) = 1$. The probability that Y_n is arbitrarily close to θ becomes 1 as n increases.

A statistic $Y_n = T(X_1, \ldots, X_n)$ is a **consistent** estimator of θ if Y_n converges in probability to θ as $n \to \infty$. In terms of notation, we say $Y_n \xrightarrow{p} \theta$ as $n \to \infty$.

For example, the sample mean $\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}$ is a consistent estimator of μ, where μ is the population mean, $\mu \equiv \mathbb{E}[X_i]$. As the sample size becomes larger, there is an increasing probability that the sample mean becomes close to μ (or there is a decreasing probability that the sample mean is different from μ). The sequence defined by $p_n = P(|X_n - \mu| \geq \epsilon)$ converges to 0 as n increases.

Therefore, the sample mean is both a consistent and an unbiased estimator of the population mean. In general, unbiasedness and consistency are two very different concepts concerning the accuracy of an estimator. An estimator can be biased but consistent, and vice versa.

For example: $\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$ is a biased but consistent estimator of the population variance. Taking the average of the first 10 observations $\frac{1}{10}\sum_{i=1}^{10}X_i$ regardless of the sample size is an unbiased but inconsistent estimator of the population mean.

A sufficient condition for Y_n to be a consistent estimator of θ is: (1) Y_n is an unbiased estimator θ, and (2) $\text{Var}(Y_n) \to 0$ when $n \to \infty$. This is an implication of the Chebychev's inequality:

$$0 \leq P(|Y_n - \mathbb{E}[Y_n]| \geq \epsilon) \leq \frac{\text{Var}(Y_n)}{\epsilon^2}$$

For example, when $\mathbb{E}[\bar{X}] = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$, it follows that $\lim_{n\to\infty}\text{Var}(\bar{X}) = 0$. As such, $\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$. This is known as the Weak Law of Large Numbers.

We also know that the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ has expectation σ^2. A bit of algebra shows that $\text{Var}(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$, where $\mu_4 = \mathbb{E}[(X-\mu)^4]$ is the fourth-central moment of X. Hence, $\text{Var}(S^2)$ also goes to zero as n increases. That is, S^2 is a consistent and unbiased estimator of the population variance.

4.2. Almost sure convergence

A sequence of random variables $X_1, X_2, \ldots,$ converges almost surely to θ if

$$P\left(\lim_{n\to\infty} X_n = \theta\right) = 1$$

In terms of notation, we say that $X_n \xrightarrow{a.s.} \theta$ as $n \to \infty$.

There is a big difference between convergence almost surely and convergence in probability.

$P\left(\lim_{n\to\infty} X_n = \theta\right) = 1$ is equivalent to: *there exists a finite number N such that for all $n > N$, the event $(|X_n - \theta| < \epsilon)$ occurs with probability 1 (surely), for any arbitrarily small $\epsilon > 0$.*

On the other hand, convergence in probability states that $\lim_{n\to\infty} P(|X_n-\theta| < \epsilon) = 1$, which is equivalent to: *there exists a finite number N such that for all $n > N$, $P(|X_n - \theta| < \epsilon) > 1 - \delta$, for any arbitrarily small $\epsilon > 0$ and $\delta > 0$.*

In almost-sure convergence, the event $|X_n-\theta| < \epsilon$ is sure (with probability 1), while for convergence in probability, the event $|X_n - \theta| < \epsilon$ occurs with probability close to one but never exactly one. Strong Law of Large Numbers: the sample mean converges almost surely to the population mean as $n \to \infty$.

Convergence almost surely is stronger than convergence in probability. The former implies the latter. But convergence in probability does not imply almost sure convergence. For example, consider the random variable $X_n = 1$ with probability $\frac{1}{n}$, and $X_n = 0$ with probability $1 - \frac{1}{n}$. Then X_n converges to 0 in probability, but X_n does not converge to 0 almost surely.

On my Github, I have posted R codes that illustrate the difference between convergence almost surely and convergence in probability.

Note that θ can be a random variable. So that when X_n converges almost surely to X, it strictly means that $P(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = \theta(\omega)) = 1$. Recall that a random variable is a *function* from some sample space Ω to the real numbers.

4.3. Convergence in distribution

A sequence of random variables $X_1, X_2, \ldots,$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

Suppose $X_1, \ldots, X_n$ is a random sample from $U[0, 1]$. Consider the largest order statistic $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

What does $X_{(n)}$ converge to? It seems like $X_{(n)}$ would converge in probability to 1. Let's verify it.

$$\begin{aligned} P(|X_{(n)} - 1| \le \epsilon) &= P(1 - \epsilon \le X_{(n)} \le 1 + \epsilon) \\ &= P(1 - \epsilon \le X_{(n)}) \\ &= (1 - \epsilon)^n \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

However consider the random variable $n(1 - X_{(n)})$.

$$\begin{aligned} P(n(1 - X_{(n)}) \leq t) &= P(1 - t/n \leq X_{(n)}) \\ &= 1 - (1 - t/n)^n \\ &\to 1 - e^{-t} \text{ as } n \to \infty \end{aligned}$$

Therefore, $n(1 - X_{(n)})$ converges in distribution to an exponential(1) random variable, since an exponential(1) random variable has the cdf $F(x) = 1 - e^{-x}$.

Convergence in probability implies convergence in distribution. In general, convergence in distribution does not imply convergence in probability, except when the converged object is a constant.