Sampling Distribution of the Sample Mean and Variance

BIOS 6611

CU Anschutz

Week 3

Properties of Random Variables

Properties of Random Variables

To derive the sampling distributions for the sample mean and variance we need to define some properties random variables that we will encounter. Let a be a constant number and X be a r.v.:

- $E(a + X) = a + E(X)$
- $Var(a + X) = Var(X)$
- $E(aX) = aE(X)$
- $Var(aX) = a^2 Var(X)$
- $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$ for independent X_i
- $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$ for independent X_i

Sampling Distribution of $\bar{X}$

Recap of the CLT

We saw in the lecture on the Central Limit Theorem that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

If $X \sim N(\mu, \sigma^2)$ then this is always true regardless of the sample size (n).

If X is not normally distributed, then $\bar{X} \dot{\sim} N(\mu, \frac{\sigma^2}{n})$ (i.e., it is approximately normal, depending on n and the shape of the original distribution considering skewness, kurtosis, etc.).

We know the population has a true distribution with its own mean, variance, etc. Each sample will take will have its own sample statistics that vary around these true values. With enough repeated samples, we would begin to observe a distribution for the sample statistic.

$\bar{X}$ is an Unbiased Estimator for the Mean (μ)

Let $X_1, ..., X_n$ be a random sample from a population with mean μ and variance σ^2, then

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

Thus, in any given sample we "expect" the sample average to be near the true population mean.

Variance of $\bar{X}$

Let $X_1, ..., X_n$ be a random sample from a population with mean μ and variance σ^2, then

$$\begin{aligned}
V[\bar{X}] &= V\left[\sum_{i=1}^{n} \frac{X_i}{n}\right] \\
&= \frac{1}{n^2} V\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} V[X_i] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

The larger the sample size, the more precise an estimator $\bar{X}$ will be.

Standard Error of $\bar{X}$

The **standard error of the mean (SEM)** measures the precision with which $\bar{X}$ estimates μ over repeated (i.e. all possible) samples of size n from a population with underlying variance σ^2:

$$\text{SEM} = \sqrt{V[\bar{X}]} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{X}}$$

In practice, we estimate the SEM by $\frac{s}{\sqrt{n}}$, where s is the standard deviation estimated from our sample.

More generally, we use the term *standard error* to refer to the sampling standard deviation of estimators of population parameters. For example, the standard deviation of (the sampling distribution of) $\hat{p}$ is the standard error of $\hat{p}$.

Sampling Distribution of $\bar{X}$

We have now shown that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

We can see that as n increases, the variability of the sample mean will decrease.

In other words, as $n \to \infty$, $\sigma_{\bar{X}} \to 0$. If we had all n data points from the population in every sample we draw, then $\bar{X}$ would be equal to the true mean μ every time we sampled.

Example of X and $\bar{X}$

The weight of 6-year old boys is normally distributed with $\mu = 40$ lbs. and $\sigma^2 = 25$ lbs.^{2}.

What is probability that X falls between 38 and 42 lbs?

Example of X and $\bar{X}$

The weight of 6-year old boys is normally distributed with $\mu = 40$ lbs. and $\sigma^2 = 25$ lbs.^{2}.

What is probability that $\bar{X}$ falls between 38 and 42 lbs for a sample of $n = 50$?

Distribution of s^2

Distribution of the Sample Variance

Thinking in terms of repeated random sampling from the population of interest, the sample variance, s^2, will also vary from sample to sample.

Recall that the theoretical definition for the variance of X is $V[X] = \sigma^2 = E\left[(X-\mu)^2\right]$. It is the average value of $(X-\mu)^2$ over all possible samples of size n.

An intuitive estimator of σ^2 might divide the *sum of deviations squared* by the sample size n:

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

However, $E\left[\frac{(X_i-\bar{X})^2}{n}\right] = \frac{n-1}{n}\sigma^2$, and so $\frac{(X_i-\bar{X})^2}{n}$ is a *biased* estimator of σ^2.

Unbiased Estimator of the σ^2

Instead of $\frac{(X_i-\bar{X})^2}{n}$, we can show that $\frac{(X_i-\bar{X})^2}{n-1}$ is *unbiased*:

$$E[s^2] = E\left[\frac{(X_i - \bar{X})^2}{n-1}\right] = \sigma^2$$

However, $E[s] \neq \sigma$, so now s is a *biased* estimator of σ (this is because the square root isn't a simple linear transformation).

Therefore, we can now state that $E\left[\frac{s^2}{n}\right] = \frac{\sigma^2}{n}$ and is unbiased, but $E\left[\frac{s}{\sqrt{n}}\right] \neq \frac{\sigma}{\sqrt{n}}$ and is biased.

How Precisely s^2 Estimates σ^2

Measure of how precisely s^2 estimates σ^2:

$$V[s^2] = \frac{2\sigma^4}{n-1} \text{ with estimate } \frac{2s^4}{n-1}$$

$$s.d.[s^2] = SE[s^2] = \sigma^2\sqrt{\frac{2}{n-1}} \text{ with estimate } s^2\sqrt{\frac{2}{n-1}}$$

This derives from the following statement. If the individual observations in a sample are $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

In other words, the χ^2 distribution with $df = n-1$ is the *sampling distribution* for $\frac{(n-1)s^2}{\sigma^2}$.

χ^2_{n-1} Sampling Distribution of $\frac{(n-1)s^2}{\sigma^2}$

If $X \sim \chi^2_k$, then $E(X) = k$ and $V(X) = 2k$.

So, if $X_1, ..., X_n$ are IID $N(\mu, \sigma^2)$ and $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, then we can show

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = (n-1) \implies E(s^2) = \sigma^2$$

$$V\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1) \implies V(s^2) = \frac{\sigma^4}{n-1}$$

We can think of this as analogous to the standard normal distribution being the sampling distribution for the statistic $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.