DSA 8010 - sampling distributions

Random samples

Random samples

So far, we have calculated probabilities related to a single value of a random variable.

Example: $Y \sim N(10, 5^2)$. What is the $P(Y \leq 6)$?

Independent and identically distributed samples

So far, we have calculated probabilities related to a single value of a random variable.

Example: $Y \sim N(10, 5^2)$. What is the $P(Y \leq 6)$?

Now, consider drawing a sample of n independent variables from the same probability distribution. These n random variables are a random sample, denoted by $Y_1, \ldots, Y_n$ or $Y_i, \quad 1, \ldots, n$.

- We can think about probabilities regarding the random sample in the same way as we think about probabilities regarding a single value of the random variable.
- We sometimes call random samples "independent and identically distributed," or i.i.d.

Example (Bernoulli random samples)

Example: Let Y be a Bernoulli random variable with $\pi = 0.85$. Knowing the value of π, we know that $Y = 1$ is more likely than $Y = 0$.

Example (Bernoulli random samples)

Example: Let Y be a Bernoulli random variable with $\pi = 0.85$. Knowing the value of π, we know that $Y = 1$ is more likely than $Y = 0$.

Now consider a random sample of $n = 4$ Bernoulli random variables with $\pi = 0.85$. Which sample is more likely to occur?

Sample A: $Y_1 = 1, Y_2 = 1, Y_3 = 0, Y_4 = 1$
Sample B: $Y_1 = 0, Y_2 = 0, Y_3 = 0, Y_4 = 1$.

Example (Bernoulli random samples)

Using rules of probabilities for independent events, we can find the probability of sample A as

$$
\begin{aligned}
P(Y_1 = 1, Y_2 = 1, Y_3 = 0, Y_4 = 1) &= P(Y_1 = 1)P(Y_2 = 1)P(Y_3 = 0)P(Y_4 = 1) \\
&= \pi \quad \cdot \quad \pi \quad \cdot \quad (1 - \pi) \quad \cdot \quad \pi \\
&= 0.85 * 0.85 * 0.15 * 0.85 \\
&= 0.09212
\end{aligned}
$$

Example (Bernoulli random samples)

Using rules of probabilities for independent events, we can find the probability of sample A as

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1, Y_3 = 0, Y_4 = 1) &= P(Y_1 = 1)P(Y_2 = 1)P(Y_3 = 0)P(Y_4 = 1) \\ &= \pi \quad \cdot \quad \pi \quad \cdot \quad (1 - \pi) \quad \cdot \quad \pi \\ &= 0.85 * 0.85 * 0.15 * 0.85 \\ &= 0.09212 \end{aligned}$$

Similarly, sample B has a probability of

$$\begin{aligned} P(Y_1 = 0, Y_2 = 0, Y_3 = 0, Y_4 = 1) &= 0.15 * 0.15 * 0.15 * 0.85 \\ &= 0.00287 \end{aligned}$$

Probability distributions of random samples

Case 1: If $Y_1, \ldots, Y_n$ are independent, random samples from a discrete probability distribution, the probability of the sample can be calculated as

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i).$$

Case 2: If $Y_1, \ldots, Y_n$ are independent, random samples from a continuous probability distribution whose density is $f(y)$, the probability of the sample can be calculated as

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i).$$

where $\prod_{i=1}^{n}$ is used to denote multiplying over the indices from 1 to n.

Sampling distributions

Sampling distribution

A sampling distribution is the probability distribution of a random sample or of a random statistic calculated from a random sample.

Sampling distribution

When considering samples of random variables, the statistics calculated from the samples are also random variables.

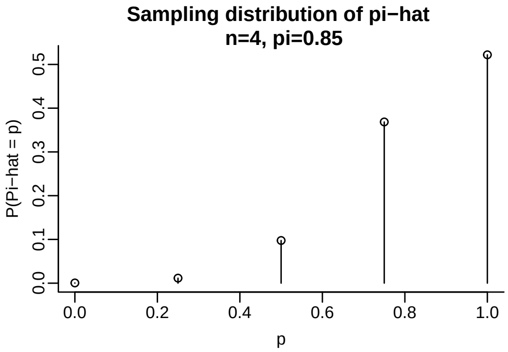
Example: in the 4 Bernoulli R.V.s, the sample proportion is one statistic that could be calculated. This is denoted by $\widehat{\pi}$ ("pi hat") and is calculated as

$$\widehat{\pi} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\text{no. successes}}{\text{no. trials}}.$$

- Different random samples will result in different values of $\widehat{\pi}$.
- The sampling distribution of $\widehat{\pi}$ gives the probability that $\widehat{\pi}$ will take on different values.

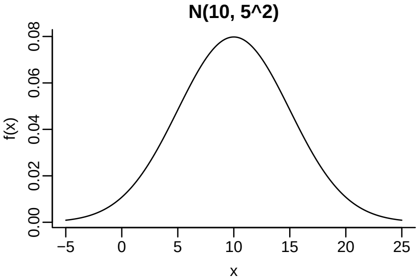
Example (Bernoulli random sample)

Here is the sampling distribution of $\widehat{\pi}$ in the sequence of 4 Bernoulli trials with $\pi = 0.85$.



Example (normal random sample)

Let $Y_1, \ldots, Y_n$ be a random sample from a $N(10, 5^2)$ distribution.



Example (normal random sample)

Let $Y_1, \ldots, Y_n$ be a random sample from a $N(10, 5^2)$ distribution.

Let $n = 5$ and take the average of $Y_1, \ldots, Y_5$. A natural statistic to summarize this sample is the sample mean,

$$\bar{Y} = \frac{\sum_{i=1}^{5} Y_i}{n}.$$

Which is more probable: $\bar{Y} > 12$ or $\bar{Y} \leq 7$?

Sampling distribution of the sample mean

Sampling distribution of the sample mean

Sampling distribution of $\bar{Y}$

Let $Y_1, \ldots, Y_n$ be a random sample from a probability distribution with $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

Let $\bar{Y}$ denote the sample mean of those n samples ($\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$).

- E($\bar{Y}$) is equal to μ.
- Var($\bar{Y}$) is equal to σ^2/n.
- Standard error($\bar{Y}$) $= \sqrt{Var(\bar{Y})}$ is equal to $\sigma/\sqrt{n}$.

Central limit theorem. As $n \to \infty$, the probability distribution of $\bar{Y}$ becomes approximately Normal(μ, $(\sigma/\sqrt{n})^2$).

Sampling distribution of the sample mean

Sampling distribution of $\bar{X}$

Let $Y_1, \ldots, Y_n$ be a random sample from a probability distribution with $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

Let $\bar{Y}$ denote the sample mean of those n samples ($\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$).

- E($\bar{Y}$) is equal to μ.

  $\rightarrow$ the most "typical" value of $\bar{Y}$ is μ.

- Var($\bar{Y}$) is equal to σ^2/n.

  $\bar{Y}$ becomes less variable as the sample size grows large.

- Standard error($\bar{Y}$) $= \sqrt{Var(\bar{Y})}$ is equal to $\sigma/\sqrt{n}$.

Central limit theorem. As $n \rightarrow \infty$, the probability distribution of $\bar{Y}$ becomes approximately Normal(μ, $(\sigma/\sqrt{n})^2$).

Sampling distribution of the sample mean

Sampling distribution of $\bar{X}$

Let $Y_1, \ldots, Y_n$ be a random sample from a probability distribution with $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

Let $\bar{Y}$ denote the sample mean of those n samples ($\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$).

- E($\bar{Y}$) is equal to μ.
- Var($\bar{Y}$) is equal to σ^2/n.
- Standard error($\bar{Y}$) $= \sqrt{Var(\bar{Y})}$ is equal to $\sigma/\sqrt{n}$.

Central limit theorem. As $n \to \infty$, the probability distribution of $\bar{Y}$ becomes approximately Normal(μ, $(\sigma/\sqrt{n})^2$).

Even if the random sample is not from a normal distribution, $\bar{Y}$ has a distribution that is approximately normal when the sample size is not too small.

Standard error

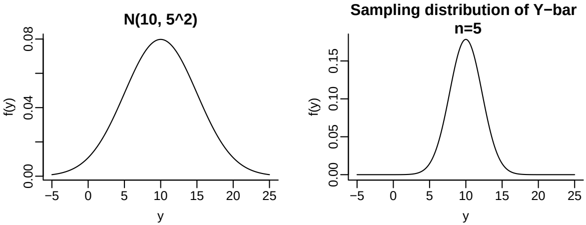
The standard deviation of a random statistic is called the standard error of the statistic.

- The standard error measures expected variability among statistics calculated from a random sample.
- The standard error gets smaller if a larger sample is taken (the statistic becomes more precise).

Standard error of $\bar{Y}$

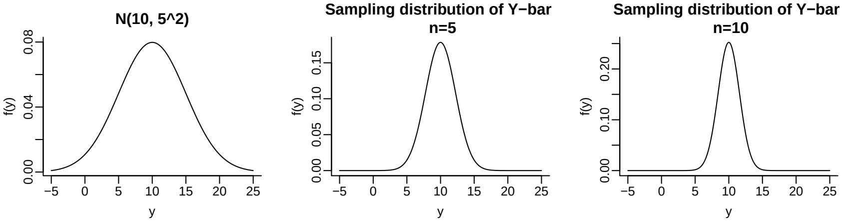
Since $Var(\bar{Y}) = \sigma^2/n$, the standard error of $\bar{Y}$ is $\sigma/\sqrt{n}$.

Sampling distribution of Y



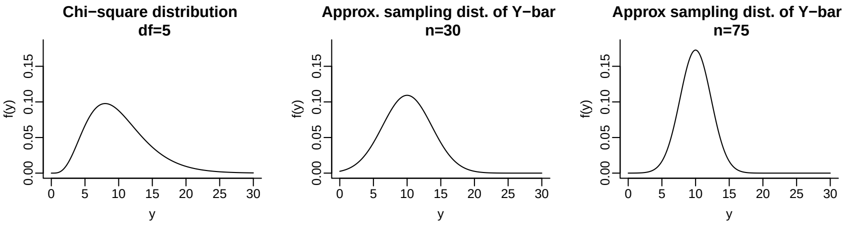
- The sampling distribution of $\bar{Y}$ is centered at μ, but is less variable than the distribution of Y.

Sampling distribution of $\bar{Y}$



- The sampling distribution of $\bar{Y}$ is different for different sample sizes.
- The standard error is smaller when n is larger.

Sampling distribution of $\bar{Y}$



- Even if the distribution of the $Y_i, \quad i = 1, \ldots, n$ is not normal, the normal distribution approximates the sampling distribution of $\bar{Y}$ (when n is large-ish, say greater than 30-40).