Toward Statistical Inference

Sampling Distributions & Central Limit Theorem (CLT)

Handout: “Summary: Sampling Distribution” and recap of last week's sampling activity

Handout: “Toward the Central Limit Theorem”

Before continuing with the actual Central Limit Theorem, let’s look at two important properties any statistic should have.

Note, in general we will refer to a statistic that is used to estimate an unknown population parameter as a so-called “estimator” (i.e. statistic = estimator)

Sampling Distribution
The sampling distribution of a statistic (e.g. the sample mean \( \bar{x} \)) is the distribution of all possible values taken by the statistic in all possible samples of the same size from the same population.

We know that our \( \bar{x} \)- value is one of the \( \bar{x} \)-values described by the sampling distribution

Bias
Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is said to be unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

Example: \( \bar{x} \) will be unbiased if the mean of the sampling distribution of \( \bar{x} \) is \( \mu \) (which in fact it is)
Toward Statistical Inference

So, if a statistic (an estimator) is unbiased, e.g. \( \bar{x} \), then it holds that

- the mean of the sample statistic is always equal to the population parameter (e.g. the mean of \( \bar{x} \) is \( \mu \))
- In repeated sampling \( \bar{x} \) will sometimes fall above the true value and sometimes fall below. However, there is no systematic tendency to over- or underestimate the parameter \( \mu \)

\[ \Rightarrow \bar{x} \text{ "is correct on average"} \]
- How close the value of the sample statistic falls to the parameter in most samples is determined by the overall spread of the sampling distribution. If individual observations have a standard deviation \( \sigma \), then sample means \( \bar{x} \) for samples of size \( n \) have standard deviation of \( \sigma / \sqrt{n} \).

\[ \Rightarrow \text{averages vary less than individual observations!} \]

Toward Statistical Inference

Recall what we learned about the sampling distribution of \( \bar{x} \):

If all possible random samples of size \( n \) are taken from some population with mean \( \mu \) and standard deviation \( \sigma \), then the sampling distribution of the sample mean \( \bar{x} \) will

- have a mean \( \mu_{\bar{x}} \) equal to \( \mu \) – the population mean
- have a standard deviation \( \sigma_{\bar{x}} \) equal to \( \frac{\sigma}{\sqrt{n}} \)

What about the shape of the sampling distribution?
Toward Statistical Inference

⇒ the shape resembles more and more the shape of a normal distribution as the sample size \(n\) increases.

**WHY?**

Toward Statistical Inference

Some guidelines regarding the necessary sample size:

- if population is already **normal**
  
  \[ \bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \]

  regardless of sample size — CLT **does not** apply!!

- if population is **symmetric and bell-shaped** resembling somewhat a normal distribution
  
  \[ \bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \]

  if sample size is \( \geq 15 \) — CLT **does** apply!!

Toward Statistical Inference

**CENTRAL LIMIT THEOREM**

The Central Limit Theorem is one of the most important Theorems in Statistics. It allows us to use normal calculations to answer questions about sample means even when the population distribution is not normal as long as the number of observations used to compute the sample mean is sufficiently large.

**Central Limit Theorem (CLT)**

If we draw a simple random sample of size \(n\) from any population with mean \(\mu\) and standard deviation \(\sigma\) and \(n\) is sufficiently large, then the sampling distribution of the sample mean \(\bar{x}\) is approximately normal:

\[ \bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \]

Toward Statistical Inference

- if population is **far from normal** (e.g. skewed or multi-modal)
  \[ \bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \]
  if sample size is \( \geq 30 \) — CLT **does** apply!!

The heavier a distribution is skewed or in the situation of very extreme observations (outliers) in the population itself, the more observations are necessary for the CLT to apply and sometimes \(n = 30\) may not be sufficient.
Example 1: According to chance magazine (1993, Vol6, Nr. 3, p.5) the mean healthy body temperature is around 98.2°F (μ) with a standard deviation of σ = 0.6. The distribution of the body temperature is known to be bell-shaped. Suppose we take a random sample of 16 adults.

- What proportion of humans has a temperature at or above the presumed norm of 98.6°F?

- What proportion of samples of size 16 have a mean temperature at or above the presumed norm of 98.6°F?

Example 2: A bottling company uses a machine to fill bottles with Cola. The bottles are supposed to contain 300 ml. In fact, the contents vary according to a normal distribution with mean μ = 298 and standard deviation σ = 3ml.

- What proportion of individual bottles contains less than 295 ml?

- What proportion of 6-packs contains less than 295?

- Did we need the CLT to derive our answers in (b)?

Application of the Law of Large Numbers: Assume that for auto accidents in the state of Iowa, the average damage (loss) is $2252 per accident.

- If you are in an accident, does $2252 apply?

- If you are in five accidents, does $2252 apply?

- If you are an insurance company and 206 of your clients have accidents, does $2252 apply?
“The law of averages.” The baseball player Tony Gwynn got a hit about 34% of the time over his 20-year career. After he failed to hit safely in six straight at-bats, the TV commentator said, “Tony is due for a hit by the law of averages.”

Is that right? Why?