Lecture 2: Sampling Distributions

BMI 713
October 24, 2017
Peter J Park

Estimation of parameters

- So far, we have assumed that the values of the parameters of a probability distributions are known.
- In the real world, these parameters are generally unknown.
- Let's use the observations in a sample to estimate a population parameter.
- **Point estimation** - calculates a single number to estimate the population parameter, e.g., *p* for a binomial distribution, μ for a normal distribution.
- **Interval estimation** - specifies a range of values for a parameter

Estimating the population mean

- We would like to estimate the mean height for graduate students at HMS
- How do we estimate the population mean μ?
- The obvious approach would be to use the mean of the sample $\bar{x}$ to estimate the unknown population mean μ
- $\bar{x}$ is called an estimator of the parameter μ
- For an unbiased estimate, the sample we have the properties of a random sample: **i.i.d. - independent and identically distributed.**
- The sample that you have is not the only that could have been selected—it is one of many possible samples

- A second sample of *n* observations could be chosen and its sample mean calculated.
- $\bar{x}_1$ and $\bar{x}_2$ are not likely to be equal (sample variability)
- Before we use $\bar{x}$ as an estimator of μ, we need to understand its properties.
- If we were to continue selecting samples of size *n* indefinitely and computing their means, we would end up with a set of values consisting entirely of samples means.
- The sample mean $\bar{x}$ is a random variable, with outcomes $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$
- How does $\bar{x}$ behave?

Sampling distribution

- The probability distribution of all possible sample means is the **sampling distribution** of the mean.
- Understanding the properties of a theoretical sampling distribution of means makes it possible to draw conclusions based on a single such sample
- It can be shown that the average of the sample means based on repeated samples of size *n* approaches the population mean μ as the number of samples selected gets large

$$E(\bar{x}) = \mu$$

- We would expect the sample means to cluster around μ
- We would also expect that the larger the sample size *n*, the more reliable the estimator $\bar{x}$

$$
\begin{aligned}
\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] &= \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] &= \frac{1}{n^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(x_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

- As *n* gets larger, $Var(\bar{x})$ decreases
- There is less variability among the sample means $\bar{x}$ than there is among the individual observations x

Standard error

- The standard deviation of $\bar{x}$ is $\sigma/\sqrt{n}$
- This is called the **standard error** of the mean

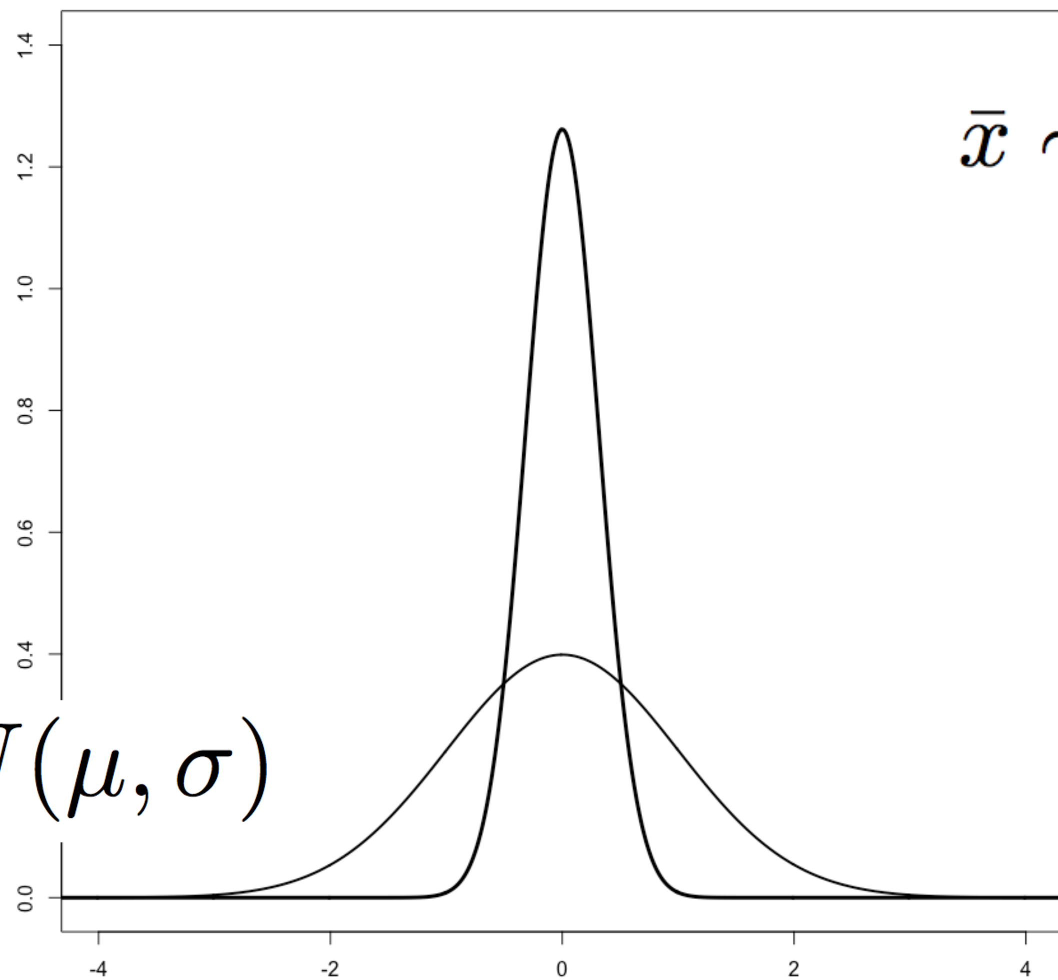
$$\bar{x} \sim \mathrm{N}(\mu, \sigma^2/n)$$

- Note that σ^2/n is determined by both the sample size and the degree of variability among the individual observations
- In general, σ quantifies the amount of variability among individuals in a population, while $\sigma/\sqrt{n}$ quantifies the variability among means of repeated samples drawn from that population

Width of the sampling distribution

sampling distribution

$$\bar{x} \sim N(\mu, \sigma^2/n)$$



population distribution

$$x_i \sim N(\mu, \sigma)$$

How does the width of the distribution scale?

The Central Limit Theorem

- If x_i is normally distributed, it can be shown that $\bar{x}$ is also normally distributed
- What if x_i is **not** normally distributed?

$$\bar{x} \sim \mathrm{N}(\mu, \sigma^2/n)$$

- Provided that *n* is large enough, **the shape of the sampling distribution is still approximately normal!**
- This is called **the central limit theorem**
- No matter what the underlying distribution of values looks like, inferences about the mean can be based on the normal distribution

What value of *n* is 'large enough' for CLT?

- It depends on the underlying distribution
- If the distribution is itself normal, then $n=1$ is large enough
- The further the population is from being normally distributed, the larger n has to be

Example

- The distribution of serum cholesterol levels for all 20- to 70-year-old males living in the United States has mean μ = 211 mg/100 ml and standard deviation σ = 46 mg/100 ml
- If a sample of size 25 is selected from this population, what is the probability that the sample has a mean of 230 or above?

- What mean value of serum cholesterol level cuts off the lower 10 of the sampling distribution?

Confidence intervals

- In reality, it is not known whether $\bar{x}$ is close to μ or not
- How do we measure the variability of the sample mean?
- Intuitively, if the variability of $\bar{x}$ is large, it is possible that the sample mean could be far from μ
- Recall

$$\bar{x} \sim \mathrm{N}(\mu, \sigma^2/n)$$

- If *n* is large,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal r.v.

- For the standard normal distribution,

$$P(-1.96 \leq z \leq 1.96) \quad = \quad 0.95$$

- Substituting $(\bar{x} - \mu)/(\sigma/\sqrt{n})$ for z

$$P\left(-1.96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \quad = \quad 0.95$$

$$P\left(-1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96\,\frac{\sigma}{\sqrt{n}}\right) \quad = \quad 0.95$$

$$P\left(\mu - 1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\,\frac{\sigma}{\sqrt{n}}\right) \quad = \quad 0.95$$

- So 95% of all sample means will lie in the interval

$$\left(\mu - 1.96\,\frac{\sigma}{\sqrt{n}}\,,\,\mu + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$$

- The statement provides information about the behavior of $\bar{x}$ if the population mean μ is known
- The problem is that μ is not known — it is what we are trying to estimate
- Back to $\mathrm{P}\left(-1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96\,\frac{\sigma}{\sqrt{n}}\right) \;=\; 0.95$

$$\mathrm{P}\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right) \;=\; 0.95$$

- The probability that the true population mean μ will be contained in

$$\left(\bar{x} - 1.96\,\frac{\sigma}{\sqrt{n}}\,,\,\bar{x} + 1.96\,\frac{\sigma}{\sqrt{n}}\right)$$ is 95%