Section_07

April 25, 2022

1 Random sampling and Sampling distributions

1.1 Random sampling

- Random sampling is selection of n samples from a population of size N
- The population has its own parameters like μ, σ.
- Statistics of the sample
    - Sample mean $\bar{x}$

$$\bar{x} = \frac{\sum x_i}{n}$$

    - Sample Variance s^2
        * Formula

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

        * Coputationally efficient formula

$$s^2 = \frac{1}{n-1} \left(\sum x_i^2 - \frac{\left(\sum x_i \right)^2}{n} \right)$$

    - Sample median
        1. Sort the values
        2. Select the middle value (if n is even median is the average of the two middle values)
    - Sample Range

$$max - min$$

    - Sample mode: The most frequent value(s)
- How to use calculator to get the mean and std deviation here

```
[18]: from IPython.display import IFrame
      IFrame(width=560, height=315, src="https://www.youtube.com/embed/AD_e7qW_Qq0")
```

```
[18]: <IPython.lib.display.IFrame at 0x7fd44d1f78b0>
```

1.1.1 Examples

Example 1 The January 1990 issue of Arizona Trend contains a supplement describing the 12 "best" golf courses in the state. The yardages (lengths) of these courses are as follows: 6981, 7099, 6930, 6992, 7518, 7100, 6935, 7518, 7013, 6800, 7041, and 6890. Calculate the sample mean and sample standard deviation, sample median, sample range and the mode. Construct a dot diagram of the data.

Example 2 An article in the Journal of Structural Engineering (Vol. 115, 1989) describes an experiment to test the yield strength of circular tubes with caps welded to the ends. The first yields (in kN) are 96, 96, 102, 102, 102, 104, 104, 108, 126, 126, 128, 128, 140, 156, 160, 160, 164, and 170. Calculate the sample mean and sample standard deviation, sample median, sample range and the mode. Construct a dot diagram of the data.

1.2 Sampling distributions

1.2.1 Distribution of sample mean $\bar{x}$

- While repeating the sampling process the sample mean $\bar{x}$ changes

- So It's a random variable and has its own distribution

- **Central limit theory**

  It states that the distribution of the sample mean $\bar{x}$ is a normal distribution with mean (mean of sample means)
  $$\mu_{\bar{x}} = \mu$$
  and std deviation
  $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  Where μ and σ are the parameters of the population.

  We can transform it to std normal distribution:
  $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- A good visualization of central limit theory here.

```
[35]: IFrame(width=1024, height=512, src="https://seeing-theory.brown.edu/
      ↪probability-distributions/index.html")
```

```
[35]: <IPython.lib.display.IFrame at 0x7fd44c164d30>
```

1.2.2 Examples

Example 3 A synthetic fiber used in manufacturing carpet has tensile strength that is normally distributed with mean 75.5 psi and standard deviation 3.5 psi. Find the probability that a random sample of n = 49 fiber specimens will have sample mean tensile strength that exceeds 75.75 psi.

Example 4 The compressive strength of concrete is normally distributed with $\mu = 2500$ psi and $\sigma = 50$ psi. Find the probability that a random sample of $n = 5$ specimens will have a sample mean diameter that falls in the interval from 2499 psi to 2510 psi.

2 Confidence Intervals

In the previous part we learned how to guess the probability of getting a specific value of the sample mean $\bar{x}$ given the population mean μ, std deviation σ and using the central limit theory.

In the real life, the parameters of the population (μ, σ) are usually unkonwn because it's almost impossible to have all observations of the population to calculate the parameters. So it makes sense to have a sample and use the sample statistics to guess the population parameters with a degree of confidence (probability).

According to the central limit theory:

$$P\left(-Z_{\frac{\alpha}{2}} < \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} < Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Where $1 - \alpha$ is the confidence interval CI

Rearranging the equation we have

$$\bar{x} - Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

In other words we can guess the range of population mean μ from the sample mean $\bar{x}$ and be confident with a specific probability value CI

$$\mu = \bar{x} \pm E$$

Where E is the estimation error and

$$E = Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

The interval length/range $(max - min) \to 2E$

We can also get the sample size using

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \times \sigma}{E}\right)^2$$

2.0.1 Student T table

- It is an approximation of the std normal distribution table
- We use it when the sample size n is very low and/or the σ of population is unkonwn
- The value $t_{\frac{\alpha}{2}, n-1}$ is an approximation of $Z_{\frac{\alpha}{2}}$
- Therefore the range of μ is

$$\mu = \bar{x} \pm \left(t_{\frac{\alpha}{2}, n-1} \times \frac{s}{\sqrt{n}}\right)$$

where s is the std deviation of the sample

- A good visualization of the difference between t and Z distributions here

```
[34]: IFrame(width=1024, height=512, src="https://rpsychologist.com/d3/tdist/")
```

```
[34]: <IPython.lib.display.IFrame at 0x7fd44c164e80>
```

2.0.2 A large sample confidence interval for a population proportion

- Approximation of a binomial RV to a normal distribution

$$Z = \frac{x - np}{\sqrt{np(1-p)}}$$

- The sample proportion $\hat{P} = \frac{x}{n}$
- The interval of the population proportion P with a confidence interval $CI = 1 - \alpha$:

$$\hat{P} - Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} < P < \hat{P} + Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

In other words

$$P = \hat{P} \pm E$$

and

$$E = Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

We can also get the sample size using

$$n = \left(\frac{Z_{\frac{\alpha}{2}}}{E} \right)^2 \times \hat{P}(1-\hat{P})$$

Note:

We only use Z table here.

2.0.3 Examples

Example 5 A civil engineer is analyzing the compressive strength of concrete. Compressive strength is normally distributed with $\sigma^2 = 1000(psi)^2$. A random sample of 12 specimens has a mean compressive strength of x = 3250 psi.

  (a) Construct a 95% confidence interval on mean compressive strength.

  (b) Construct a 99% confidence interval on mean compressive strength. Compare the width of this confidence interval with the width of the one found in part (a).

Example 6   Dairy cows at large commercial farms often receive injections of bST (Bovine Somatotropin), a hormone used to spur milk production. Bauman et al. (Journal of Dairy Science, 1989) reported that 12 cows given bST produced an average of 28.0 kg/d of milk. Assume that the standard deviation of milk production is 2.25 kg/d.

(a) Find a 99% confidence interval for the true mean milk production.

(b) If the farms want the confidence interval to be no wider than ± 1.25 kg/d, what level of confidence would they need to use?

Example 7   A confidence interval estimate is desired for the gain in a circuit on a semiconductor device. Assume that gain is normally distributed with standard deviation $\sigma = 20$.

(a) Find a 95% CI for m when n = 10 and $\bar{x} = 1000$.

(b) Find a 95% CI for m when n = 25 and $\bar{x} = 1000$.

(c) How does the length of the CIs computed change with the changes in sample size?

(d) How large must n be if the length of the 95% CI is to be 40?

Example 8   A research engineer for a tire manufacturer is investigating tire life for a new rubber compound and has built 16 tires and tested them to end-of-life in a road test. The sample mean and standard deviation are 60139.7 and 3645.94 kilometers. Find a 95% confidence interval on mean tire life.

Example 9   The fraction of defective integrated circuits produced in a photolithography process is being studied. A random sample of 300 circuits is tested, revealing 13 defectives. Calculate a 95% CI on the fraction of defective circuits produced by this particular tool.

Example 10   Of 1000 randomly selected cases of lung cancer, 823 resulted in death within 10 years.

(a) Calculate a 95% confidence interval on the death rate from lung cancer.

(b) Using the point estimate of p obtained from the preliminary sample, what sample size is needed to be 95% confident that the error in estimating the true value of p is less than 0.03?