

BIOS 600: Principles of Statistical Inference

Sampling Distribution of the Mean

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 8

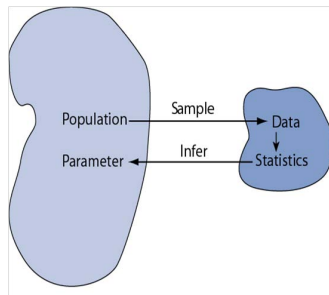
Overview

While the topic may seem a little theoretical, the sampling distribution of the mean plays a very important role in statistical inference. We need these points in order to make statistical inference and draw conclusions about the results we see.

The basic idea, called the *Central Limit Theorem*, is that for *any* distribution with a well-defined mean and variance, the distribution of the means computed for samples of size n is approximately normal. This is an extremely important result for hypothesis testing and construction of confidence intervals later in the course.

Statistical Inference

- ▶ Statistical inference is the act of generalizing from a sample to a population with an estimated degree of certainty.
- ▶ We want to know about *parameters* in the population.
- ▶ We calculate *statistics* in the sample to learn about the parameters.



Parameters and statistics

It is imperative to understand the distinction between parameters and statistics.

	Parameters	Statistics
Source	Population	Sample
Calculated?	No	Yes
Constant?	Yes (unless Bayesian)	No
Example notation	μ, σ, π	\bar{x}, s, p

Parameters and statistics



Still confused about parameters and statistics? Let's talk Plato instead. Plato presented his famous **Allegory of the Cave** in *The Republic*. It may be helpful to think of parameters as *Platonic forms* and the statistics we calculate as the shadows on the wall of the cave. Here is a short video clip describing this allegory: Allegory of the Cave Claymation.

Note: if you've never read *The Republic*, fall term may not be the ideal time to tackle it!

Sampling Distribution of the Mean

Consider the following hypothetical situation. Suppose we can list all the members of a population. We then will do the following.

- ▶ Take a random sample of size n . Call this Sample 1.
- ▶ Compute the sample average, \bar{x}_1 .
- ▶ Put the sample back, and take a second random sample also of size n , calling it Sample 2.
- ▶ Compute the sample average, \bar{x}_2 .
- ▶ Repeat this many times, creating a dataset that consists of these sample averages $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots$

What is the distribution of the statistics $\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots$?

Simulation study

Let's try it out! Roughly 60% of all part-time college students in the U.S. are female ($\pi = 0.60$). Define a variable x to take value 1 if a student is female and value 0 if a student is not female.

- ▶ What is the distribution of X ?
 - ▶ X is Bernoulli. A Bernoulli random variable has mean π and variance $\pi(1 - \pi)$ (binomial with $n = 1$)
- ▶ If we take a random sample, the average $\bar{x} = \hat{\pi}$ is an *estimate* of the proportion of females in the population of interest.
- ▶ If we take random samples of part-time college students and calculate the proportion of females in each sample, denoted $\hat{\pi}$, what values will you see? Will you get 0.60 every time?
- ▶ Let's run a simulation study and see.
- ▶ How did the variability of the sampling distribution depend on the size of our random samples?

Central Limit Theorem

The *Central Limit Theorem* says that for a population with mean μ and standard deviation σ , the important three properties of the distribution of sample averages \bar{x} hold:

- ▶ The mean of the sampling distribution is identical to the population mean μ .
- ▶ The standard deviation of the distribution of the sample averages is $\frac{\sigma}{\sqrt{n}}$, called the *standard error* of the mean.
- ▶ For n large enough (in the limit as $n \rightarrow \infty$), the shape of the sampling distribution is approximately normal (Gaussian)

Back to proportions

Going back to our example, the Central Limit Theorem tells us that the distribution of sample averages $\hat{\pi}_i$ should have mean π and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$. The standard deviation result is found by taking the standard deviation of the Bernoulli, $\sqrt{\pi(1-\pi)}$, and dividing by \sqrt{n} .

A useful rule of thumb in this setting is that for the Central Limit Theorem to kick in, we want $n\pi > 10$ and $n(1-\pi) > 10$.

What? That was a binary random variable!

The Central Limit Theorem tells us that the sample averages are normally distributed if we have enough data. This result holds even if our original variables (here, a binary variable) are not normally distributed.

Let's see what happens with a skewed distribution

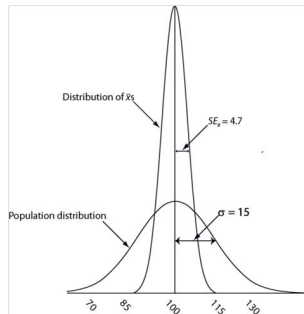
Mean birth weights

Suppose we conduct research in a population in which the mean birth weight is 3500g with a standard deviation of 500g. The birth weight distribution is approximately normal if we restrict to term births.

Distribution of average birth weights

IQ Distribution

The distribution of IQ in the general population is normal with mean 100 and standard deviation 15. Suppose we draw samples of size $n = 10$ from this population. Then from the Central Limit Theorem, we know that the distribution of the sample averages will also be normal with mean 100 and standard deviation $\frac{15}{\sqrt{10}} = 4.7$. If our data on IQ are exactly normal, then the distribution of the sample averages will also be exactly normal.



If our data on IQ are not exactly normal (for example, in BIOS 600, they may be skewed), then the rule of thumb is that at a minimum 30 observations are needed for the Central Limit Theorem to kick in.

Example: IQ in BIOS 600

Suppose that I give a random sample of size $n = 50$ of you an IQ test, and the sample average score is 120. Does this mean that BIOS 600 students are smarter than average?

Example: IQ in BIOS 600

The Central Limit Theorem tells us that the distribution of means of samples of size 50 from this population is also normal with mean $\mu = 100$ and standard error $\frac{\sigma}{\sqrt{50}} = \frac{15}{7.07} = 2.12$. Recall that $Z = \frac{\bar{X} - 100}{15}$ is a standard normal random variable, so if $\bar{x} = 120$, then $z = \frac{120 - 100}{2.12} = 9.43$. Consulting Table A.3 in the text, the probability of a z-value greater than this is extremely small (less than 0.001). What does this mean?

Example: IQ in BIOS 600

What are the upper and lower limits that enclose 95% of the means of samples of size 50 drawn from the population? Because 2.5% of the area under the standard normal curve falls above $z = 1.96$ and 2.5% of the area falls below $z = -1.96$, we know that the middle 95% falls between these two values, or in mathematical notation,

$$Pr(-1.96 \leq Z \leq 1.96) = 0.95.$$

So our limits for the standard normal distribution are $-1.96 \leq Z \leq 1.96$. We can transform back into a statement about the distribution of \bar{X} as follows.

Example: IQ in BIOS 600



Why take the middle 95%?

OK, you're right, we could have constructed an asymmetric interval that took the bottom 95% or top 95% of the data, instead of the middle 95%. Or, we could have taken the bottom 40%, the top 40%, and then the middle 15%. Why didn't we do something funky like that?

Symmetric intervals in this setting are the shortest intervals that capture the appropriate proportion of the means.

Example: IQ in BIOS 600

$$\begin{aligned}-1.96 &\leq Z \leq 1.96 \\-1.96 &\leq \frac{\bar{X}-100}{2.12} \leq 1.96 \\-1.96(2.12) &\leq \bar{X} - 100 \leq 1.96(2.12) \\-4.16 &\leq \bar{X} - 100 \leq 4.16 \\-4.16 + 100 &\leq \bar{X} \leq 4.16 + 100 \\95.84 &\leq \bar{X} \leq 104.16\end{aligned}$$

So about 95% of the sample averages in random samples of size 50 drawn from the general population would have means between 95.84 and 104.16.

Example: IQ in BIOS 600

What does this mean for our sample with average score 120? Well, one option is that the sample of BIOS 600 students comes from a population different from the general population (hmm, are BIOS 600 students normal?). Another option is that it does come from the general population, but we've encountered a fairly rare event in terms of the composition of our random sample. We can't be sure.

Example: IQ in BIOS 600

How does the interval coverage depend on the size of our random sample? Let's consider taking samples of varying size and see. (Recall $\mu = 100$ and $\sigma = 15$.)

n	$\frac{\sigma}{\sqrt{n}}$	Interval Enclosing 95% of Averages
1	15.00	$70.6 \leq \bar{X} \leq 129.4$
20	3.35	$93.4 \leq \bar{X} \leq 106.6$
50	2.12	$95.8 \leq \bar{X} \leq 104.2$
1000	0.47	$99.1 \leq \bar{X} \leq 100.9$

As the sample size increases, we expect intervals to be more narrow.

Example: IQ in BIOS 600

Finally, let's consider a more complicated question. How large would our random samples need to be for 95% of their averages to lie within ± 10 of the population mean μ ?

To solve this, find the sample size n for which

$$Pr(\mu - 10 \leq \bar{X} \leq \mu + 10) = 0.95.$$

We can do the calculation by using algebra to get this statement in terms of a z-score.

Example: IQ in BIOS 600

$$\begin{array}{ccccc} \mu - 10 & \leq & \bar{X} & \leq & \mu + 10 \\ -10 & \leq & \bar{X} - \mu & \leq & 10 \end{array}$$

$$\frac{-10}{\frac{15}{\sqrt{n}}} \leq \frac{\bar{X} - \mu}{\frac{15}{\sqrt{n}}} \leq \frac{10}{\frac{15}{\sqrt{n}}}$$

$$\frac{-10}{\frac{15}{\sqrt{n}}} \leq Z \leq \frac{10}{\frac{15}{\sqrt{n}}}$$

$$-0.67\sqrt{n} \leq Z \leq 0.67\sqrt{n}$$

Because we want 95% of our values between ± 10 , we set the upper and lower limits equal to 1.96 and -1.96, respectively, and solve.

Example: IQ in BIOS 600

We can use either the upper or lower limit (same answer). So we solve with $z = 1.96$ as

$$z = 1.96 = 0.67\sqrt{n}$$

$$2.92 = \sqrt{n}$$

$$2.92^2 = n$$

$$8.56 = n$$

Because we like to be conservative when we deal with sample size, we round up (always) so we need samples of size $n = 9$ to expect that 95% of our sample averages will lie within 10 of the population mean.

Reading for Next Time

- ▶ Pagano and Gauvreau, Chapter 9