Introductory Statistics Lectures

The Central Limit Theorem

Sampling distributions

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:49:45 2009)

Contents

1 The Central Limit Theorem

1.1 Introduction

R TIP OF THE DAY: RANDOM NUMBER GENERATORS

Normal random numbers:
`x=rnorm(n, mean=0, sd=1)`
Creates `n` random numbers from a normal distribution and stores them in `x` .

R Command

In R, random number generators have a "r" prefix.

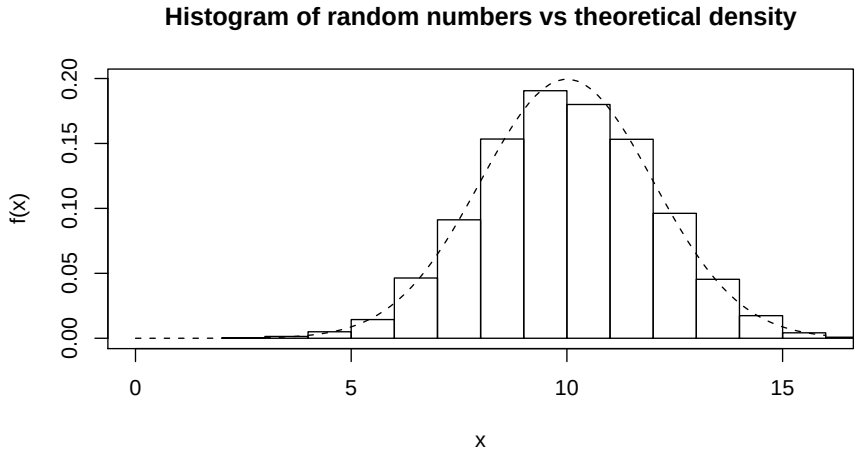
Random number generators are useful for simulating random variables and exploring complex systems using Monte Carlo methods.

Example 1. Generate 3 sets of 8 random random numbers from a normally distributed population with a mean of 10 and standard deviation of 2.

```
R: rnorm(8, mean = 10, sd = 2)
[1]  9.6562  6.9938  8.3588  8.1606  9.8177  9.5088 10.8508
[8]  6.4181
R: rnorm(8, mean = 10, sd = 2)
[1]  7.0716  6.9356  6.6494 12.9266 12.0327 10.1264 12.6311
[8] 10.2955
R: rnorm(8, mean = 10, sd = 2)
[1]  9.6340  8.3440  7.1764  9.1673  7.3618 10.7930  7.5256
[8] 12.5522
```

Comparison of histogram of 5,000 randomly generated numbers to density function.

```
R: x.rand = rnorm(5000, mean = 10, sd = 2)
R: curve(dnorm(x, mean = 10, sd = 2), 0, 16, lty = "dashed",
+      ylab = "f(x)", main = "Histogram of random numbers vs ↩
    theoretical density")
R: hist(x.rand, prob = T, add = T)
```



POPULATION MEANS AND SAMPLE MEANS

College aged women's systolic blood pressures (in mm Hg) are normally distributed with a mean of 114.8 and a standard deviation of 13.1.

Question 1. If we randomly measured 10 women in this subgroup, would the mean blood pressure be 114.8?

Question 2. If we conducted another random sample of 10 women would the mean blood pressure be the same?

Question 3. If we repeatedly sample all the possible combinations of 10 women what would the **distribution of sample means** look like?

Sample means
Take 5 random samples each with 5 women from the subgroup and computing the mean **to estimate the population mean.** $\bar{x} = \{104, 116, 119, 115, 110\}$

Question 4. Is the sample mean the same each time? Why?

Question 5. What is the range of sample mean values?

Question 6. What is a quick estimate of their standard deviation?

Question 7. How could we improve the sample mean's estimate of the population mean?

Sample means
Now let's increase n from 5 to 20 for each sample and observe the effect on the sample means. Take 5 random samples each with 20 women from the subgroup and computing the mean **to estimate the population mean.** $\bar{x} = \{114, 115, 117, 114, 115\}$

Question 8. What is the range of sample mean values?

Question 9. What is a quick estimate of their standard deviation?

Question 10. Does the variation in the $\bar{x}$'s increase or decrease if we increased the sample size?

Question 11. Does the sample mean — an estimate of the population mean — get better or worse as the sample size increases?

1.2 Sampling distributions

SAMPLING VARIABILITY. DEFINITION 1.1

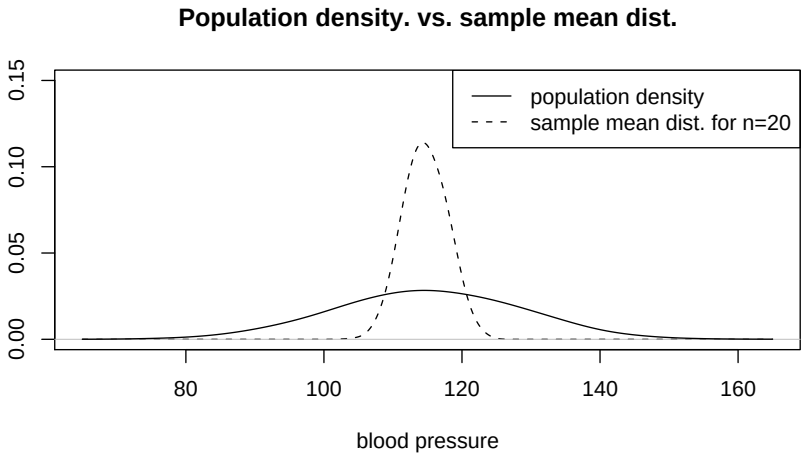
From sample to sample the value of a statistic will vary due to random fluctuations (sampling error).

DEFINITION 1.2 SAMPLING DISTRIBUTION.
A probability distribution that describes a statistic used to estimate a parameter. Result of sampling variability. Describes how precise and accurate a statistic is for measuring a population parameter.

DEFINITION 1.3 SAMPLING DISTRIBUTION OF THE MEAN.
Probability distribution of sample means $\bar{x}$ of sample size n.

DEFINITION 1.4 SAMPLING DISTRIBUTION OF THE PROPORTION.
Probability distribution of sample proportions $\hat{p}$ of sample size n.

Comparison of population dist. vs. sample mean dist.
Repeatedly sampling 20 women from the subgroup and plotting the distribution of sample means $\bar{x}$ along with the population density:



DEFINITION 1.5 UNBIASED ESTIMATORS.
If the mean of a sampling distribution for a statistic equals the population parameter it is unbiased. On average, it correctly estimates the parameter. Biased statistics tend to be wrong on average.
unbiased statistics mean, variance, proportion
biased statistics median, range, standard deviation

Good estimators are often unbiased but "biased" tends to have negative overtones and not all biased statistics are bad. Biased statistics are often used with great effectiveness (eg. standard deviation). Bias is just **one** measure of how "good" a statistic is, there are other measures such as consistency, efficiency, and sufficiency.

Question 12. Why would we use standard deviation instead of variance if it is biased?

Why sample with replacement

- Sampling with replacement results in independent events, making the probabilities easier to describe and the resulting formulas become much simpler. We will assume sampling with replacement for many cases later in this class.
- As we have seen, when our sample size is small relative to the population ($n/N \leq 0.05$), sampling without replacement can be approximated as with replacement.

1.3 Central Limit Theorem: sampling dist. of $\bar{x}$

CENTRAL LIMIT THEOREM (CLT). DEFINITION 1.6

The **sampling distribution of** $\bar{x}$ with random sample size n will be **normally distributed if**

1. the population is normally distributed **or**
2. the sample size $n > 30$. [1]

and the mean and standard deviation of the sampling distribution will be:

$$\boxed{\mu_{\bar{x}} = \mu} \tag{1}$$

$$\boxed{\text{standard error: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}} \tag{2}$$

where μ and σ are the population mean and standard deviation. VERY IMPORTANT!

What the CLT tells us

- If we satisfy the CLT, then $\bar{x}$ has a normal distribution and a known mean and standard deviation.
- $\sigma_{\bar{x}}$ **tells us how precisely we can estimate** μ **using** $\bar{x}$**.**
- Finally, we can **characterize** $\bar{x}$**'s sampling error!**
- Increase n to increase the accuracy of the estimate for μ. (increasing n decreases $\sigma_{\bar{x}}$).

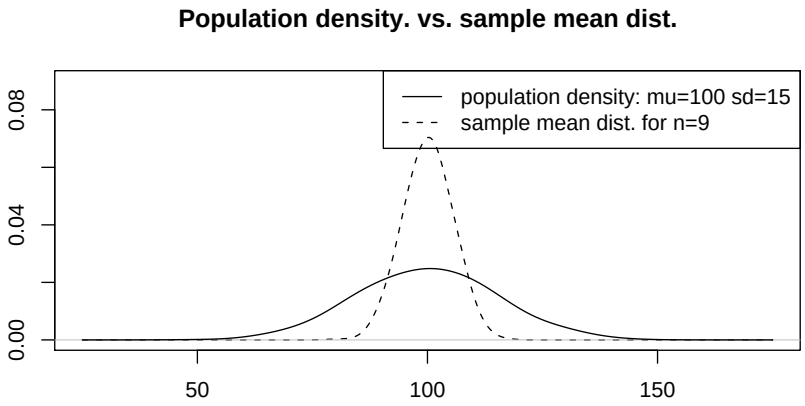
Beyond sampling distributions, CLT says that variables determined by complex systems often have a normal distribution (the convolution of a number of density functions tends to the normal). That's one reason why we see it frequently in nature: height, weight, ...

Question 13. If we repeatedly sample all the possible combinations of 10 women what would the **distribution of sample means** look like?

Case I when CLT applies.

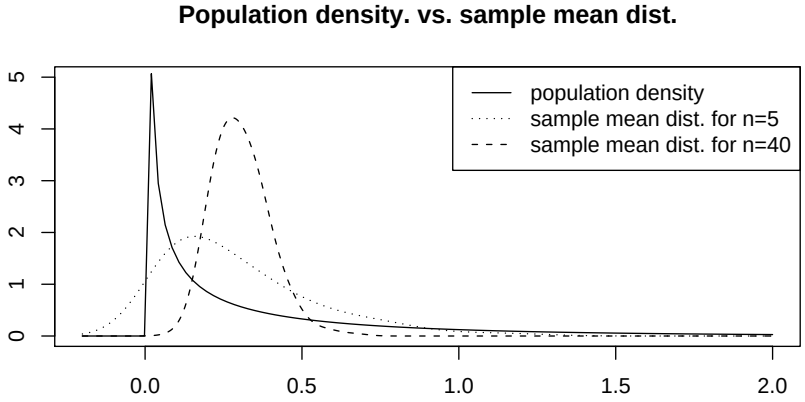
[1] Will be **approximately** normal.

Since x is normally distributed we see the sample mean $\bar{x}$ distribution is also normally distributed as predicted by the CLT.

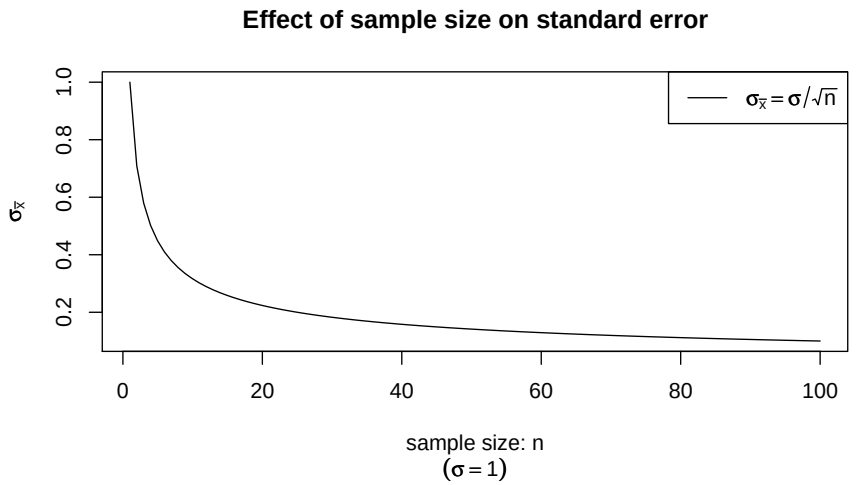


The Empirical rule tells us:
95% of data x : $\mu_x \pm 2\sigma_x = 100 \pm 2 \cdot 15 = 100 \pm 30$
95% of sample means $\bar{x}$: $\mu_{\bar{x}} \pm 2\sigma_{\bar{x}} = 100 \pm 2\frac{15}{\sqrt{9}} = 100 \pm 10$

Case II when CLT applies.
If x is **not** normally distributed (it's a gamma dist.) we see the sample mean $\bar{x}$ distribution is **not** normally distributed when $n \not> 30$, but is approximately normally distributed when $n > 30$ as predicted by the CLT.



Relationship of $\sigma_{\bar{x}}$ to n



Increasing n decreases the variation in the sample means $\sigma_{\bar{x}}$. Thus, you can increase the precision of $\bar{x}$ as an estimate for μ by increasing n. However, each time you double the precision (halve $\sigma_{\bar{x}}$) you must increase the sample size by fourfold.

Tips for solving problems
First write down what is given: μ, σ, n, ..., then if the question is about:

1. an **individual**: use μ and σ.
2. the **mean** or a **number of individuals**: use $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$.

1.4 Summary

- Sampling distribution: describes distribution of a statistic.
- Smaller sample distribution variation indicates we can more precisely estimate the population parameter.
- CLT: $\bar{x}$ normally distributed if
  - x normally distributed or
  - $n > 30$

  then:

$$\boxed{\mu_{\bar{x}} = \mu}$$

$$\boxed{\text{standard error: } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}}$$

1.5 Additional Examples

Men's hip breadths are normally distributed with a mean of 14.4 in and a standard deviation of 1.0 in.

Question 14. If one man is randomly selected, find the probability that his hip breadth is between 15 and 17 in.

Question 15. If 25 men are randomly selected, find the probability that their mean hip breadth is between 15 and 17 in.

Question 16. You need to build a bench that will seat 18 male football players. What is the minimum length of the bench if you want a 0.975 probability that it will fit all 18 men.

Question 17. What's wrong with our answer from the previous question?