

Lab 9: Normal approximations for means
STT 421: Summer, 2004
Vince Melfi

In previous labs where we investigated the distribution of the sample mean and sample proportion, we often noticed that the distribution is approximately symmetric and bell-shaped. This is no accident: For “large” sample sizes, the normal distribution provides a good approximation for the distributions of the sample mean and sample proportion. For the sample proportion, we saw this by comparing the CDF of the binomial distribution (which controls the CDF of the sample proportion) to the CDF of a normal distribution. In this lab we will investigate this approximation for the sample mean, in the process learning about **if** statements in SAS. We’ll again use the dataset **various**, and will compare the simulated sampling distribution of the mean and proportion with the normal approximation. (Throughout the lab we’ll treat this simulated sampling distribution as if it is the true sampling distribution.) For reference below, here is the output of **proc means** applied to the **various** data.

Variable	N	Mean	Std Dev	Minimum	Maximum
exp	10000	0.9905400	0.9925513	0.000599137	8.1489949
cau	10000	1.6701817	79.8181789	-1657.96	5608.86
nor	10000	7.0342084	4.0260695	-8.4390507	23.2139770
bin	10000	0.3021000	0.4591913	0	1.0000000

Normal populations

The variable **nor** is supposed to contain a normally distributed population. Of course, no finite population can be exactly normally distributed, but as the histogram of **nor** will suggest, its distribution is very close to normal.

From Section 5.2 of the text, we know that if we take a random sample of size n from a population that is normally distributed with mean μ and standard deviation σ , then the sample mean \bar{x} is exactly normally distributed with mean μ and standard deviation σ/\sqrt{n} . In our case we have $\mu \approx 7$ and $\sigma \approx 4$, so we expect that if we take a sample of size 4, \bar{x} will have a normal distribution with mean 7 and standard deviation $4/\sqrt{4} = 2$. First we simulate the sampling distribution as before and look at its histogram.

```
data various;
  infile 'u:\msu\course\stt\421\summer04\various.dat';
  input exp cau nor bin;

proc surveysselect data = various n = 4 rep = 2500 out = norsamp4;
  id nor;

proc univariate data = norsamp4 noprint;
  output out = normeans4 mean = Mean;
  var nor;
  by replicate;
```

```

proc univariate data = various;
  var nor;
  histogram nor;
  title 'Normal population';

proc univariate data = normeans4;
  var mean;
  histogram mean;
  title 'Sample mean, n=4, normal population';

run;

```

This program should produce a histogram of the population and a histogram of the sample means. Both should be centered around 7, with the sample mean histogram having a smaller spread. Look at the histograms to make sure this is true.

Now we want to compute some probabilities for the sampling distribution and the normal distribution to make sure they are reasonably close. We'll look at

$$P(\bar{X} \leq 4); \quad P(\bar{X} \leq 11); \quad \text{and} \quad P(7 \leq \bar{X} \leq 10).$$

We'll use **if** statements to construct **dummy** variables that will help us to find the probabilities for the sampling distribution.

```

data norprobs1;
  set normeans4;
  IF mean <= 4 THEN lessthan4 = 1;
  ELSE lessthan4 = 0;
  IF mean <= 11 THEN lessthan11 = 1;
  ELSE lessthan11 = 0;
  IF mean >= 7 AND mean <= 10 THEN between7and10 = 1;
  ELSE between7and10 = 0;
  drop mean replicate;

proc freq data = norprobs1;
  title 'probabilities for distribution of sample mean, n=4,
  normal population';

run;

```

Explanation

1. We create a new dataset called **norprobs1**. The statement **set normeans4** indicates that we want to start with the dataset **normeans4**.
2. The **IF** statements create the dummy variables. For example, the first **IF ELSE** pair tells SAS to create a variable called **lessthan4** and set it equal to 1 whenever the

sample mean is less than or equal to 4, and set it equal to 0 whenever the sample mean is greater than 4.

3. The **drop mean replicate** statement tells SAS to remove the variables **mean** and **replicate** from the **norprobs1** dataset. (They were part of the **normeans4** dataset that we started with.)
4. Now the **proc freq** statement gives us what we need. The relative frequency of 1 gives the probability. For example, the relative frequency of 1 for the variable **lessthan4** gives $P(\bar{X} \leq 4)$.

Now we'll compute the same probabilities for a normal distribution with mean 7 and standard deviation 2.

```
data blah;
  x1 = cdf('normal', 4, 7, 2);
  x2 = cdf('normal', 11, 7, 2);
  x3 = cdf('normal', 10, 7, 2) -
        cdf('normal', 7, 7, 2);
```

```
proc print data = blah;
```

```
run;
```

Questions

1. What is $P(\bar{X} \leq 4)$? What is the normal approximation to this probability?
2. What is $P(\bar{X} \leq 11)$? What is the normal approximation to this probability?
3. What is $P(7 \leq \bar{X} \leq 10)$? What is the normal approximation to this probability?

Skewed population

The population contained in the variable **exp** is skewed, as you can see by looking at its histogram. For skewed populations, it takes a larger sample size for the normal approximation to be accurate.

We'll repeat the above computations for the **exp** population. We'll have to change some of the numbers to reflect the fact that the mean and standard deviation of this population are both approximately 1, but the basic idea is the same.

```
proc surveysselect data = various n = 4 rep = 2500 out = expsamp4;
  id exp;
```

```
proc univariate data = expsamp4 noprint;
  output out = expmeans4 mean = Mean;
  var exp;
```

```

    by replicate;

proc univariate data = various;
    var exp;
    histogram exp;
    title 'Skewed (exp) population';

proc univariate data = expmeans4;
    var mean;
    histogram mean;
    title 'Sample mean, n=4, skewed (exp) population';

data expprobs1;
    set expmeans4;
    IF mean <= 0.4 THEN lessthan4tenths = 1;
    ELSE lessthan4tenths = 0;
    IF mean <= 2 THEN lessthan2 = 1;
    ELSE lessthan2 = 0;
    IF mean >= 1 AND mean <= 2 THEN between1and2 = 1;
    ELSE between1and2 = 0;
    drop mean replicate;

proc freq data = expprobs1;
    title 'probabilities for distribution of sample mean, n=4,
        skewed (exp) population';

data blah;
    x1 = cdf('normal', 0.4, 1, 0.5);
    x2 = cdf('normal', 2, 1, 0.5);
    x3 = cdf('normal', 2, 1, 0.5) -
        cdf('normal', 1, 1, 0.5);

proc print data = blah;

run;

```

Questions

1. Compare the probability that \bar{X} is less than 0.4 to the normal approximation.
2. Compare the probability that \bar{X} is less than 2 to the normal approximation.
3. Compare the probability that \bar{X} is between than 1 and 2 to the normal approximation.

You likely found that the normal distribution did not perform very well as an approximation this time. That is due to the fact that the population distribution is skewed. In such cases,

a larger sample size is needed before the sample mean's distribution is closely approximated by the normal distribution.

Increasing the sample size

Repeat the same sort of simulation as above using the `exp` population, but this time with $n = 25$ rather than $n = 4$. How does the normal distribution perform as an approximation to the distribution of \bar{X} ? In particular, answer the following questions (remember that you need to compute the correct mean and standard deviation for the normal distribution).

1. What is $P(\bar{X} \leq .9)$? What is the normal approximation for this probability?
2. What is $P(\bar{X} \leq 1.2)$? What is the normal approximation for this probability?
3. What is $P(0.85 \leq \bar{X} \leq 1.1)$? What is the normal approximation for this probability?