

MATH 243 PARTIAL LECTURE NOTES (25 APRIL
2008)

N. CHRISTOPHER PHILLIPS

Suppose we have a random variable which takes real numbers as values (quantitative, not categorical). Recall that it has a *distribution*.

For a continuous distribution, the probability of a value of this random variable being in an interval is the area under a suitable part of a curve. Example: Heights of young adult women, measured in inches, (approximately) have the normal distribution $N(64, 2.7)$. However, in the following the distribution need not be continuous. Example: The number of accidents on a given day at the intersection of Belt Line Road and Delta Highway.

For each sample size n , there is a *sampling distribution* for the mean \bar{x} of simple random samples of size n .

Example:

The distribution of heights of young adult women tells you the probability that, if you choose a young adult woman at random, her height is, say, in the range $61.3 \leq x \leq 66.7$. (This range is the heights within one standard deviation of the mean.) The sampling distribution for samples of size 9 tells you the probability that, if you choose 9 young adult women at random (a simple random sample of size 9), their mean height is, say, in the range $61.3 \leq \bar{x} \leq 66.7$.

What do you expect?

What we expect:

The simple random sample is more likely to include some short women and some tall women than to contain only short women or only tall women. (Having only tall women is like flipping a coin 9 times and having it come up tails every time.) So the spread of the distribution should be less.

Date: 25 April 2008.

The mean of the sampling distribution is the average of sample averages. Therefore, one expects the mean of the sampling distribution to be the same as that of the original distribution.

Facts about sampling distributions:

Suppose we are given a probability distribution. Consider sampling distribution for the mean \bar{x} of simple random samples of size n from the given distribution.

- If the original distribution has mean μ , then so does the sampling distribution. (The statistic \bar{x} is an *unbiased* estimator of μ .)
- If the original distribution has standard deviation σ , then the sampling distribution has standard deviation $\frac{\sigma}{\sqrt{n}}$. (So the spread is smaller.) We often call this the *sampling standard deviation*.
- If the original distribution is normal, then so is the sampling distribution.
- If the original distribution is not normal, *and the population size is much larger than the sample size*, then the sampling distribution of \bar{x} is closer to normal than the original distribution, and in fact can be better approximated by a normal distribution for larger sample size n .

This last fact is a consequence of the *Central Limit Theorem*. See pages 280–285 of the book. There are three sets of pictures which show this theorem in action, on pages 282, 283, and 285.

Example: Heights of young adult women, measured in inches, (approximately) have the normal distribution $N(64, 2.7)$.

So the sampling distribution of mean heights for simple random samples of size 9 is approximately $N(64, 0.9)$.

Example: Order processing times (in minutes) at Wang’s Widgets Inc. have a highly nonnormal distribution, skewed strongly to the right, with mean $\mu = 7$ and standard deviation $\sigma = 21$.

Assume that Wang’s Widgets Inc. processes millions of orders (at least).

Then the sampling distribution of mean ages for simple random samples of size 10,000 is approximately $N(7, 0.21)$. Note that it is approximately normal, even though the original distribution is not.

Examples of computations:

Heights of young adult women, measured in inches, (approximately) have the normal distribution $N(64, 2.7)$.

Choose one young adult woman at random. What is the probability that her height is greater than 64 inches? Greater than 66.7 inches?

Choose a simple random sample of 4 young adult women. What is the probability that the mean height of the 4 women in the sample is greater than 64 inches? Greater than 66.7 inches?

Choose a simple random sample of 9 young adult women. What is the probability that the mean height of the 9 women in the sample is greater than 64 inches? Greater than 66.7 inches?

Choose a simple random sample of 10,000 young adult women. What is the probability that the mean height of the 10,000 women in the sample is greater than 64.054 inches? Greater than 639.46 inches?

Choose one young adult woman at random. What is the probability that her height is greater than 64 inches?

Answer: About $\frac{1}{2}$. In an exactly normal distribution, the probability of being above the mean is exactly $\frac{1}{2}$. (Here the distribution is only approximately normal.)

Greater than 66.7 inches?

66.7 is one standard deviation above the mean. By the Rule of Thumb, the probability of being more than one standard deviation away from the mean is about $1 - 0.68 = 0.32$. The probability of being more than one standard deviation above the mean is half this, or 0.16.

Choose a simple random sample of 4 young adult women. What is the probability that the mean height of the 4 women in the sample is greater than 64 inches?

The sampling distribution is approximately $N(64, 2.7/\sqrt{4}) = N(64, 1.35)$. So the answer is about $\frac{1}{2}$, just as before.

Greater than 66.7 inches?

66.7 is now *two* sampling standard deviations above the mean. By the Rule of Thumb, the probability of being more than two standard deviations away from the mean is about $1 - 0.95 = 0.05$. The probability of being more than two standard deviations above the mean is half this, or 0.025. (From Table A, one can get the better approximation 0.0228.)

Choose a simple random sample of 9 young adult women. What is the probability that the mean height of the 9 women in the sample is greater than 64 inches?

The sampling distribution is approximately $N(64, 2.7/\sqrt{9}) = N(64, 0.9)$. So the answer is about $\frac{1}{2}$, just as before.

Greater than 66.7 inches?

66.7 is now *three* sampling standard deviations above the mean. By the Rule of Thumb, the probability of being more than three standard deviations away from the mean is about $1 - 0.997 = 0.003$. The probability of being more than three standard deviations above the mean is half this, or 0.0015. (From Table A, one can get the better approximation 0.0013.)

Choose a simple random sample of 10,000 young adult women. What is the probability that the mean height of the 10,000 women in the sample is greater than 64.054 inches?

The sampling distribution is approximately $N(64, 2.7/\sqrt{10,000}) = N(64, 0.027)$.

64.054 is now *two* sampling standard deviations above the mean. By the Rule of Thumb, the probability of being more than two standard deviations away from the mean is about $1 - 0.95 = 0.05$. The probability of being more than two standard deviations above the mean is half this, or 0.025. (From Table A, one can get the better approximation 0.0228.)

Greater than 63.946 inches?

Similar reasoning gives: About 0.9772 (Table A) (about 0.9975 according to the Rule of Thumb).

Note that, with probability about 95%, the sample mean \bar{x} is in the very small interval $63.946 \leq \bar{x} \leq 64.054$.

Why it is important:

Suppose we know that the population distribution is approximately normal, with $\sigma = 2.7$, but we do not know μ . Suppose we choose a simple random sample of size 10,000, and we find that, for this particular sample, $\bar{x} = 64.0317$. Then we believe, with 95% confidence, that

$$\mu - 2(\sigma/\sqrt{10,000}) \leq \bar{x} \leq \mu + 2(\sigma/\sqrt{10,000}),$$

equivalently,

$$\bar{x} - 2(\sigma/\sqrt{10,000}) \leq \mu \leq \bar{x} + 2(\sigma/\sqrt{10,000})$$

(\bar{x} is within the distance $2(\sigma/\sqrt{10,000})$ of μ), so

$$63.9777 \leq \mu \leq 64.0857.$$

95% confidence: We got this estimate by a method that gives the right answer 95% of the time.

I didn't ask for the probability that the mean height of a simple random sample of 10,000 young adult women is at greater than 66.7. This probability turns out to be about $7.6 \cdot 10^{-24}$.

More examples of computations:

Order processing times (in minutes) at Wang's Widgets Inc. have a highly nonnormal distribution, skewed strongly to the right, with mean $\mu = 7$ and standard deviation $\sigma = 21$.

Choose one order at random. What is the probability that its processing time is greater than 7 minutes? Greater than 28 minutes?

Choose a simple random sample of 4 orders. What is the probability that the mean of the processing times of the 4 orders in the sample is greater than 7 minutes? Greater than 28 minutes?

Choose a simple random sample of 9 order processing times. What is the probability that the mean of the processing times of the 9 orders in the sample is greater than 7 minutes? Greater than 28 minutes?

Choose one order at random. What is the probability that its processing time is greater than 7 minutes? Greater than 28 minutes?

We don't know, since we don't know the distribution. (If it were normal, the probability would be about 0.5 for the first question, and

0.16 for the second. For typical distributions of the sort, the probability of \bar{x} being greater than the mean is probably less than $\frac{1}{2}$, and the probability of \bar{x} being greater than the mean by at least one standard deviation might well be larger than 0.16.)

Compare: What is the probability that its processing time is less than -14 minutes (that is, more than one standard deviation below the mean)?

Choose a simple random sample of 4 or 9 orders. What is the probability that the mean of the processing times of the 4 or 9 orders in the sample is greater than 7 minutes? Greater than 28 minutes?

Again, we can't tell. The sampling distributions are not close to normal. (If they were, we would, for example, find a probability of 0.0013 that the mean of the processing times of the orders in the samples of size 9 is less than -14 .)

Choose a simple random sample of 10,000 order processing times. What is the probability that the mean of the processing times of the 10,000 orders in the sample is greater than 7.42 minutes? Greater than 6.56 minutes?

Now the sampling distribution is approximately normal, in fact, approximately $N(7, 21/\sqrt{10,000}) = N(7, 0.21)$. We can proceed as for heights of young adult women.

Since 7.42 is two sampling standard deviations above the sampling mean, and 6.56 is two sampling standard deviations above the sampling mean, the answers are the same as before: about 0.0228 and about 0.9972 (following Table A).

The Central Limit Theorem revisited:

Recall that the Central Limit Theorem implies that, for sufficiently large sample sizes (but much smaller than the population size), the sampling distribution of \bar{x} becomes approximately normal.

Restated: Consider the mean \bar{x} of a sufficiently large sample, of size n , each member of which is drawn independently from a distribution with mean $\mu = 0$ and standard deviation σ . Then the sampling distribution of \bar{x} is approximately normal with mean 0 and standard deviation $\frac{\sigma}{\sqrt{n}}$.

In fact, the sample mean \bar{x} will still be approximately normally distributed,

- even if the original distributions are all different,
- even if the original distributions are far from normal (in different ways),
- even if the standard deviations are not all the same (as long as they don't vary too much),
- even if there are small violations of the independence condition.

Thus, if a random variable is the result of combining many small unknown effects not strongly related to each other, its distribution is expected to be roughly normal.

This is why the following are roughly normally distributed:

- Math SAT scores.
- Heights of adult men.
- In many cases, the results of repeated measurement of the same thing with the same scientific instrument.
- Many other examples.

In particular, the requirement we will often see, that the data be approximately normally distributed, is not as unreasonable as you might have thought when this issue first arose.