Electronics and
Computer Science

UNIVERSITY OF
Southampton

COMP6235
The Normal Distribution, Estimation and Confidence Intervals

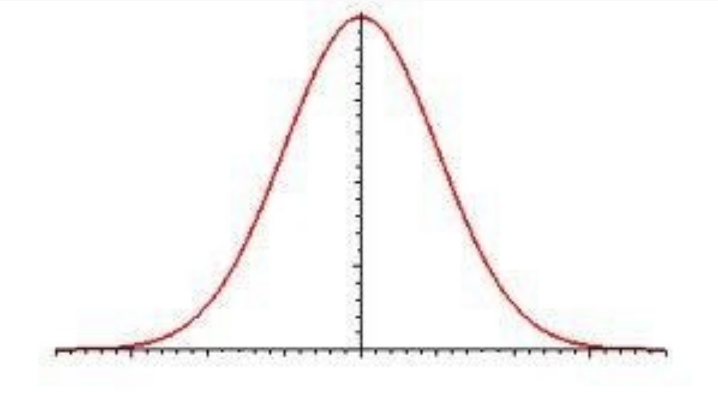
Dr. Jian Shi

Jian.Shi@soton.ac.uk

The normal distribution

- The normal distribution is the classic "bell curve".
- We've seen that we can produce one by adding or averaging a large-enough group of random variates from *any* distribution.
- It can also be specified as a probability density function.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



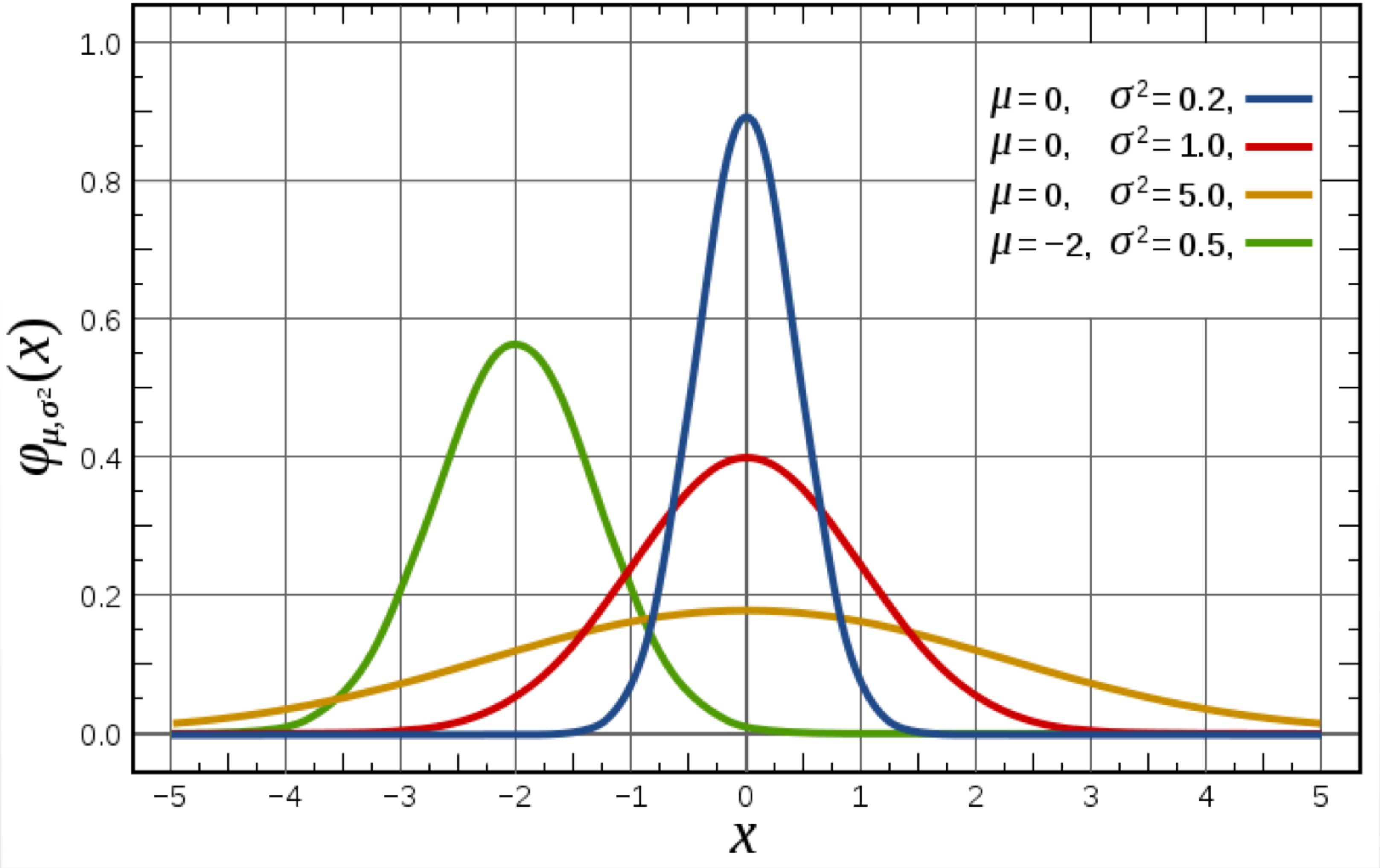
The normal distribution

- The normal distribution is central to statistical inference.

- A particular normal distribution is fully characterized by just two parameters: the mean, μ, and the standard deviation, σ.

- In other words, once you've said where the centre of the distribution is, and how wide it is, you've said all you can about it. The general shape of the curve is consistent.

The normal distribution



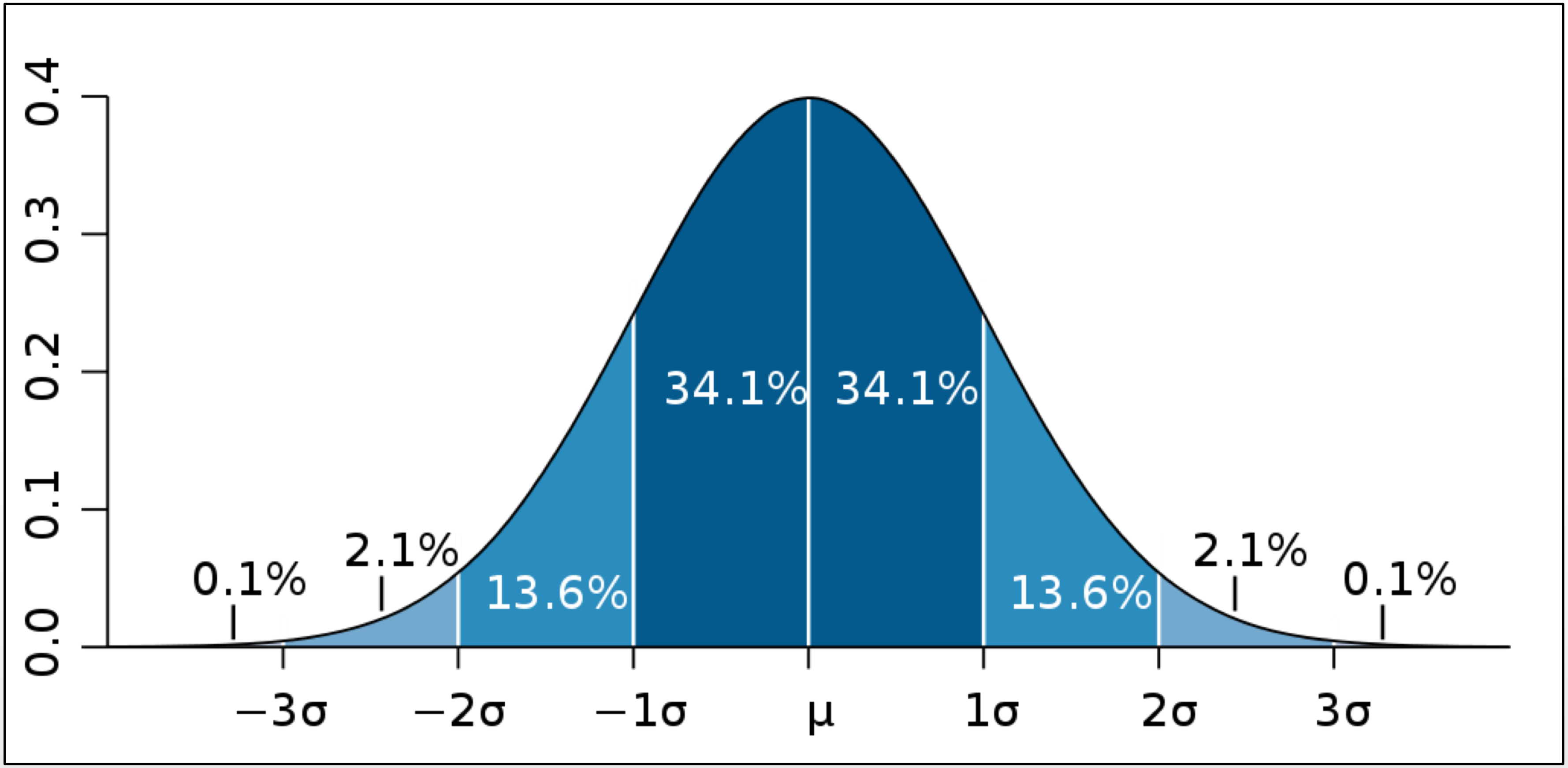
Standard normal distribution

- Because the normal distribution has this constant shape we can translate any instance of it into a standardized form. This is because of the scaling relations we discussed when we discussed the proof of the CLT.
- For each value, subtract the mean and divide by the standard deviation.
- This gives us the standard normal distribution which has $\mu = 0$ and $\sigma = 1$ (red line on previous slide).
- Values on the standard normal indicate the number of standard deviations that an observation is above or below the mean. They're also called **z-scores**.

Areas under the normal curve

- The normal distribution's consistent shape is useful because we can say precise things about areas under the curve.
- It's a probability distribution so the area sums to 100%.
- 68% of the time the variate will be within plus or minus one standard deviation of the mean (i.e., a z-score between -1 and 1).
- 95% of variates will be within two standard deviations.
- 99.7% of variates will be within three standard deviations.

Areas under the normal curve



Variates from the normal distribution

- Suppose we have a normal distribution with a mean of 100 and a standard deviation of 10.
- We can reason about how unusual particular values are.
- For example, only 0.1% of cases will have a score higher than 130.
- Around 95% of cases will lie between 80 and 120. Conversely, only about 5% of cases will be more than 20 points away from 100.
- 34% of cases will be between 100 and 110.

Z-tables

- In practice these days we use statistical calculators like Python to figure out these areas under the normal curve. One often uses area under the curve from -infinity to a particular value of z
- In the past, you had to look up a pre-computed "Z-table".

Positive Z-score	Area remaining under the curve to the right of this point
1.0	0.1587
1.5	0.0668
2.0	0.0228
2.5	0.0062

- You can also use the Python-function norm.cdf() or norm.ppf()

Notable z-scores

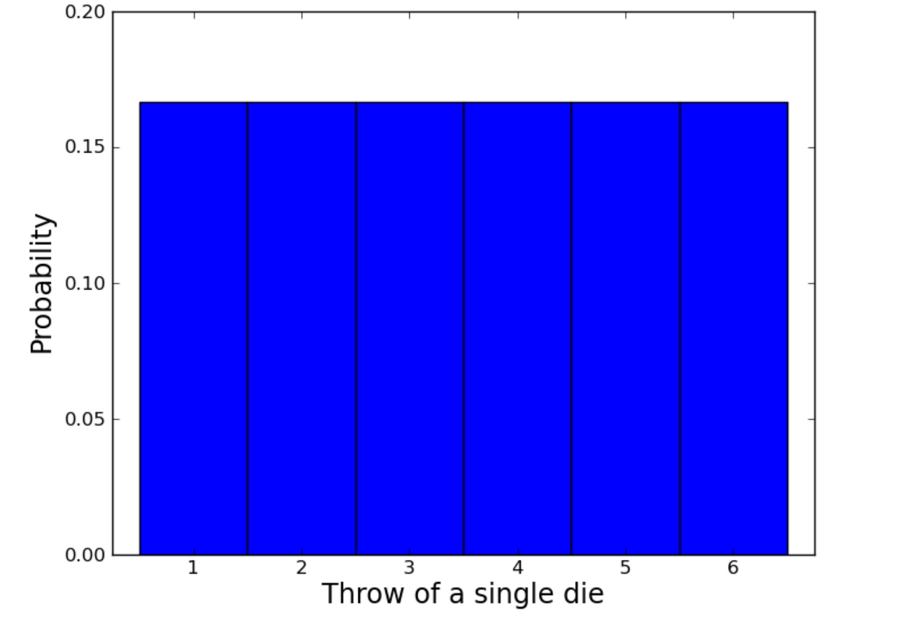
A few useful z-scores to remember...

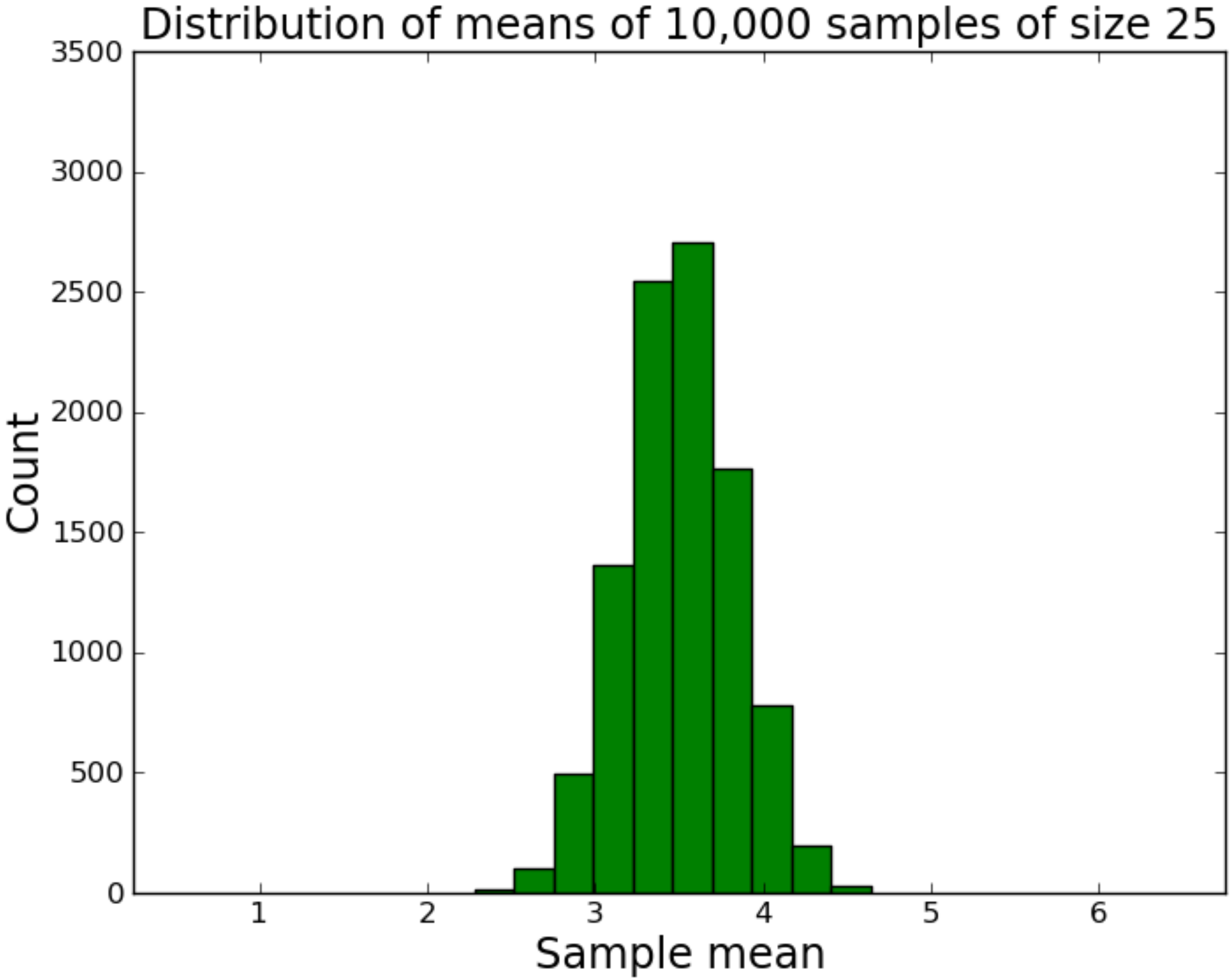
- Z = 1.645 leaves 5% of the curve to the right.
- Z = 1.96 leaves 2.5% of the curve to the right.
- Z = 2.58 leaves 0.5% of the curve to the right.

Thinking of a particular sample mean as a variate from a normal distribution

- Recall the uniform distribution of integers between 1 and 6 we get from throwing a single die.

- We found previously that if we repeatedly take samples of size N from that distribution, we end up with our collection of sample means being approximately normally distributed.




Distribution of means of 10,000 samples of size 25
Count
Sample mean
0
500
1000
1500
2000
2500
3000
3500
1
2
3
4
5
6

Sampling distribution of the mean

- What can we say about this approximately normal distribution of sample means?
- The mean is the same as the original population mean, i.e., 3.5 in this case.
- The standard deviation is the same as the original distribution's, scaled by 1 / sqrt(N), where N is the sample size.
- In the N=25 case, that's 1.708 / sqrt(25) = 0.342.

A note about why we're doing this

- Remember that in real cases nobody is interested in the green histogram for its own sake.

- If you really had the resources to collect 10,000 samples of size 25, you'd just call it one huge sample of 250,000.

- In the real case you get only your single sample of size 25, and you're trying to make inferences about the population based on just that.

- These computational experiments where we generate many such samples are attempts to stand back from that one-shot perspective.

A problem

- So it looks as though we're ready to draw conclusions about how unusual certain sample means might be.

- After all, we've got an approximately normal distribution (of sample means) and we know ***its*** mean and sd.

- But there's a problem: we're helping ourselves to information we wouldn't have in the real case.

- The original population's mean and standard deviation are typically exactly the things we're trying to find, not pre-given information.

Sampling distribution of the variance

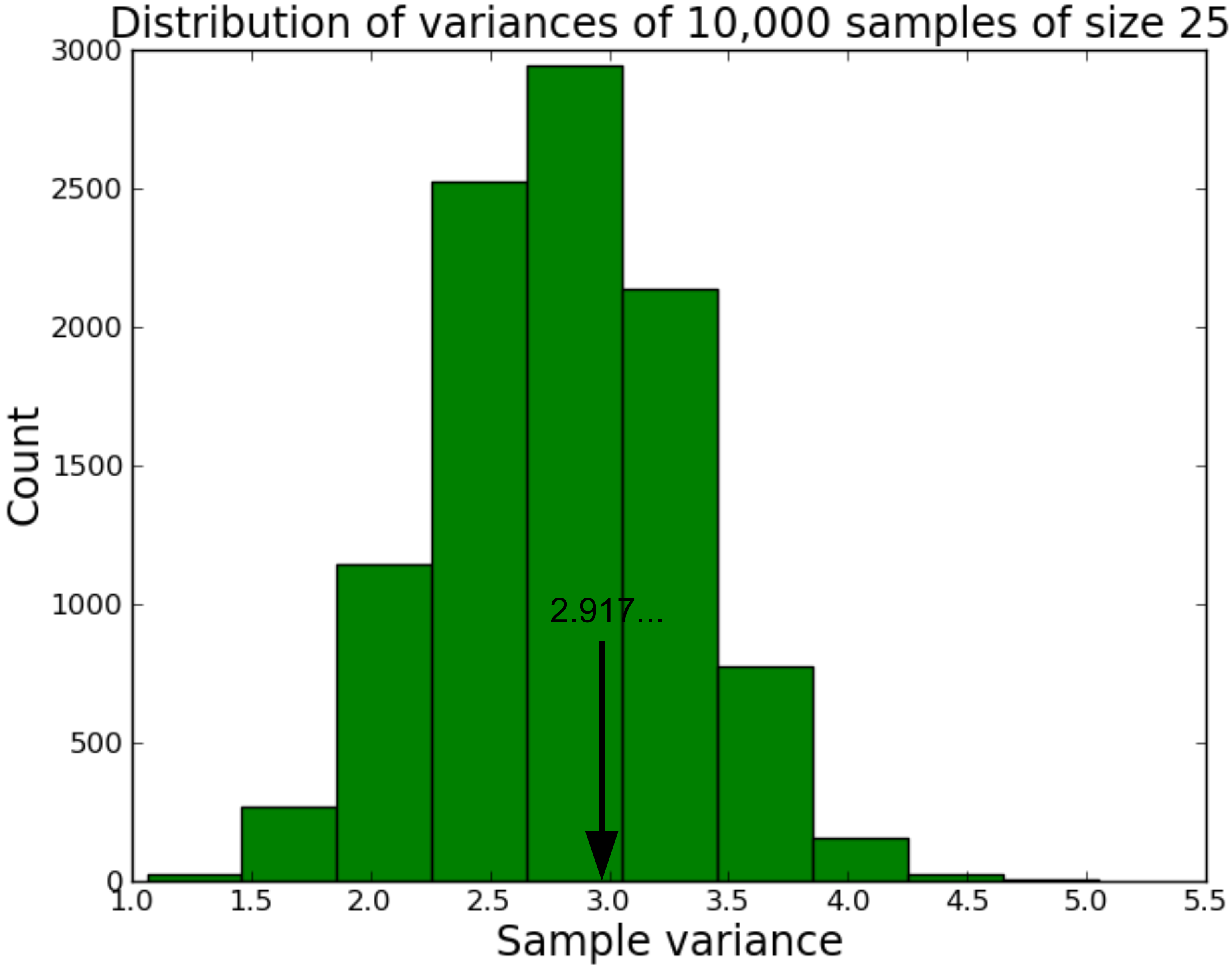
- We need to work with the only things we have, i.e., the properties of our ***sample***.

- We know that the mean of our sample is a "good guess" for the mean of the population.

- What about the variance of our sample? We haven't systematically checked the relationship between sample variance and population variance yet.

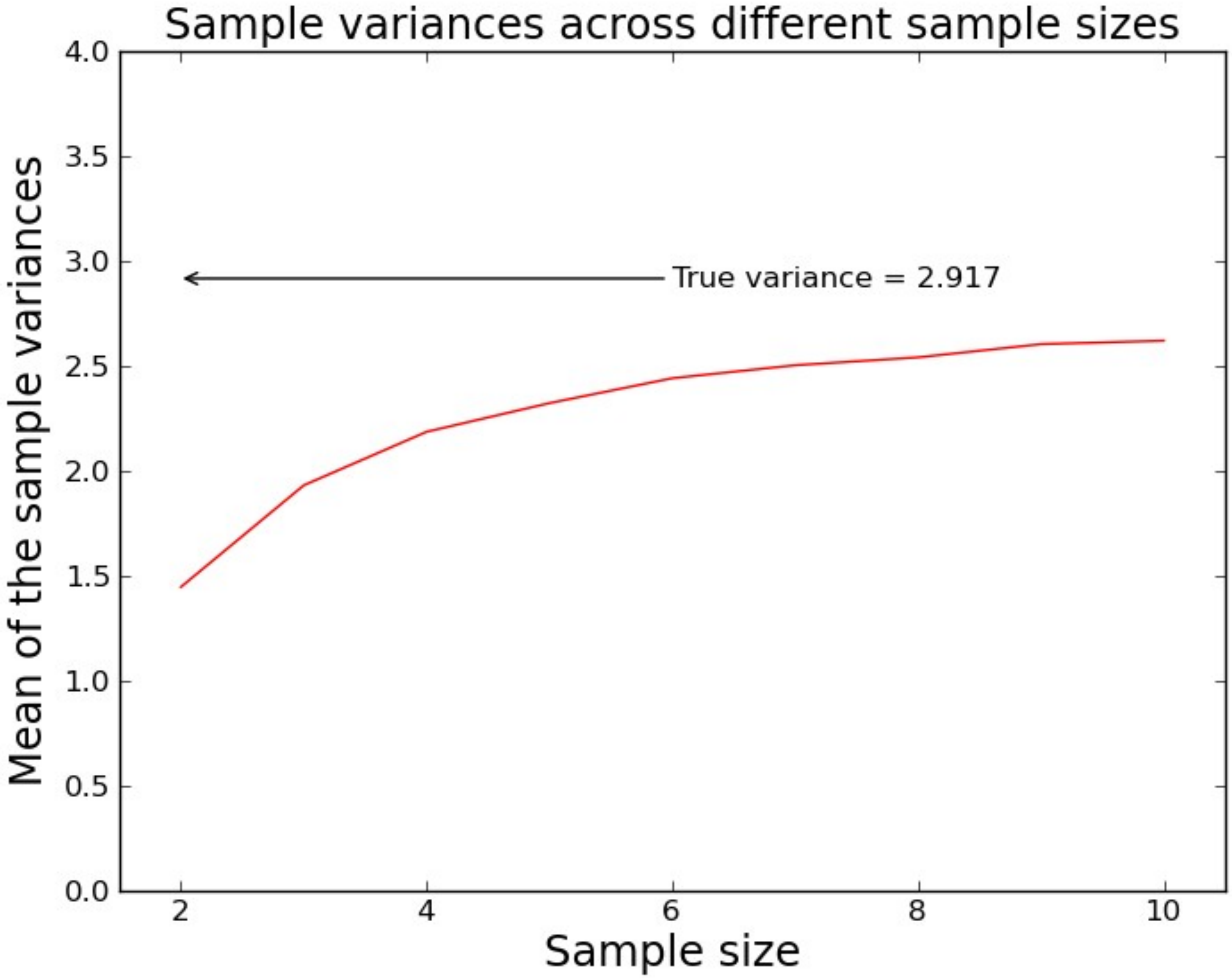
- Let's take 10,000 samples of size 25, calculate the variance in each case, and consider the distribution of those sample variances.

E
C


Distribution of variances of 10,000 samples of size 25
Count
3000
2500
2000
1500
1000
500
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
Sample variance
2.917...

Sampling distribution of the variance

- At first glance this all looks good.
- The variances of many samples of size 25 seem to be themselves roughly normally distributed and they seem to be zeroing in on the true value of 2.917.
- But let's look more closely: for sample sizes between 2 and 10, we'll try collecting 10,000 samples and noting the average value of the calculated sample variance.
- Turns out there's a systematic problem of underestimation.


Sample variances across different sample sizes
Mean of the sample variances
Sample size
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
2
4
6
8
10
True variance = 2.917

Biased and unbiased estimators

- The sample mean is an *unbiased estimator* of the population mean.
- This means that although our sample mean may be quite far from the true value, it's equally likely to be high or low.
- The sample variance is a *biased estimator* of the population variance.
- The sample variance will systematically underestimate the population variance, especially so for small sample sizes.

The sample variance and sample SD

- Bessel's correction is needed in order to find an unbiased estimator of the population variance.

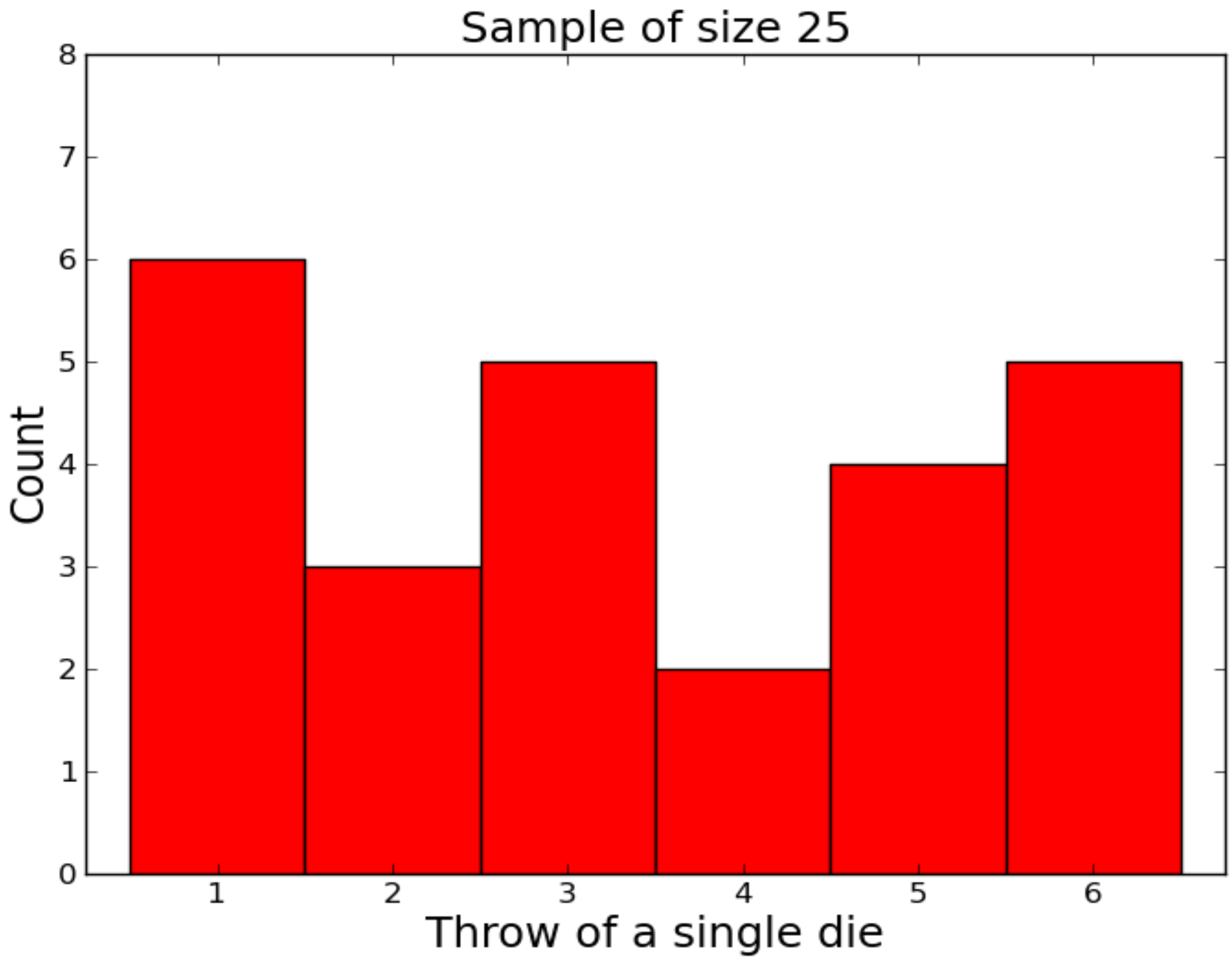
- This means simply that we need to divide through by (N - 1) instead of N when calculating the variance and standard deviation.

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

- The underestimation happens because we're using the same small set of numbers to estimate both the mean and the variance.

A realistic case

- At last we're in a position to take a particular sample of size 25 and see how we could use it to reason about the population.
- For the sake of the exercise, we'll pretend that we don't already know the true values of the population mean and variance.


Sample of size 25
Count
0
1
2
3
4
5
6
7
8
1
2
3
4
5
6
Throw of a single die

A realistic case

Some output from Python...

```
Mean of the sample is 3.4
Variance of the sample is 3.44
Sample variance, estimating pop variance, is  3.58333333333
SD of the sample is 1.8547236991
Sample SD, estimating pop SD, is 1.8929694486
```

So, based on our sample information, the best guess for the population mean is 3.4, and for the population standard deviation it's 1.893. (Not bad guesses: true values are 3.5 and 1.708.)

A realistic case

- We can place this information in a wider context.

- We know that our sample mean is "noisy", i.e., it's really drawn from an approximately normal distribution of possible sample means.

- Our best guess for ***that*** distribution is that its mean is 3.4 and its standard deviation is 1.893 / sqrt(25) = 0.379.

- That calculation gives us the ***standard error of the mean***, i.e., the estimated standard deviation of the sampling distribution for the mean.

Confidence intervals

- If the sampling distribution of the mean is normally distributed, we can say something about how unlikely it is to get an extreme value from that distribution.
- We know, for example, that getting a z-score more extreme than ±3 only happens 0.3% of the time.
- Using our best estimates for the sampling distribution's mean and SD, that's the equivalent of saying that sample means outside the range of 3.4 ± (3 x 0.379) will only happen 0.3% of the time.

Confidence intervals

- A very handy z-score is 1.96, because it leaves exactly 2.5% of the distribution to the right.

- If we consider both edges of the normal distribution, that means that only 5% of the time will we get values more extreme than $z = \pm1.96$.

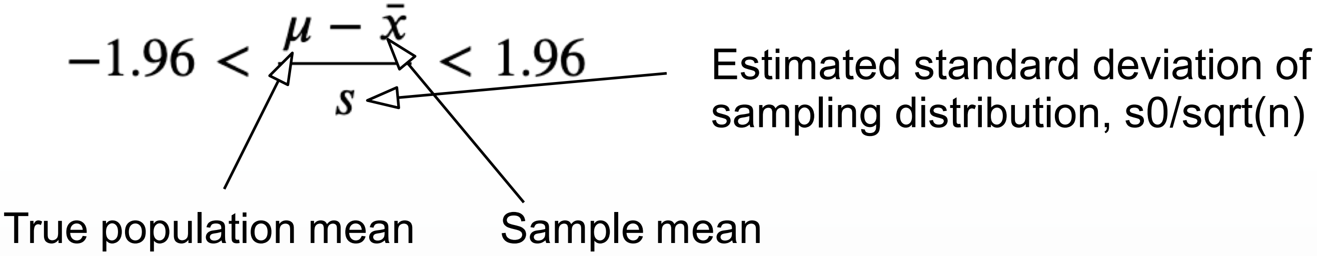
- So, 95% of the time, we expect our sample mean to lie within the range 3.4 ± (1.96 x 0.379). That's between 2.657 and 4.143.

Confidence intervals

- Remember that 3.4 is our absolute best guess for the mean. (We don't know that the true value is 3.5.)
- But we also know that 3.4 is unlikely to be exactly right. We know that we're vulnerable to sampling error.
- If the mean of a particular ***sample*** is within ±1.96 standard errors of the population mean 95% of the time, we can also reverse this logic.
- We can conclude that 95% of the time, the true ***population mean*** is within ±1.96 standard errors of our sample mean.

Confidence intervals

- To see this in formulae:
  - The Z-table tells us that with 95% probability we have



$$-1.96 < \frac{\mu - \bar{x}}{s} < 1.96$$

Estimated standard deviation of sampling distribution, s0/sqrt(n)

True population mean

Sample mean

  - Hence, with 95% probability we have:

$$-1.96s + \bar{x} < \mu < 1.96s + \bar{x}$$

i.e. with 95% probability the true mean is in

$$\bar{x} \pm 1.96s$$

Confidence intervals

- And that's all a confidence interval is.
- In this case, we would say that the 95% confidence interval for the true population mean is between 2.657 and 4.143.
- Note that the right answer, 3.5, is within that interval.
- Not guaranteed: 5% of the time, the real value will be outside the interval. Of course we won't know when!
- Confidence intervals can be calculated for different confidence levels (90%, 99%) with different z-scores, and can be calculated for quantities other than the mean.

Additional material

- David M. Lane's tutorials on the normal distribution and on sampling.
- A guide to reporting standard deviations and standard errors.