Using s^2 and s as Estimators for σ^2 and σ

Recall the definition of the sample variance:

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

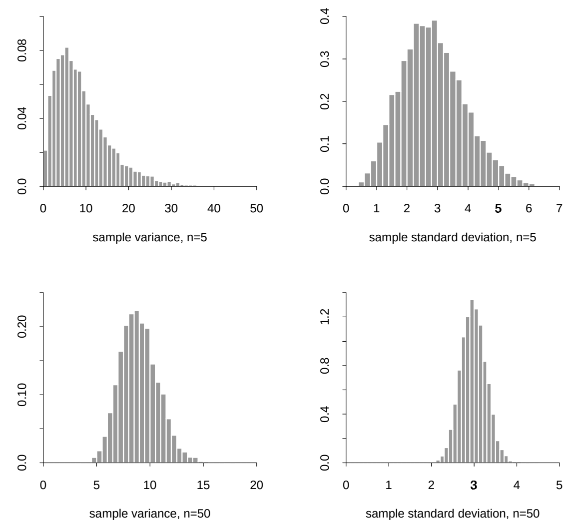
This is a statistic, computed from the sample, $x_1, \ldots, x_n$.

We would like to know whether s^2 is a good estimator of σ^2, and also whether s is a good estimator of σ.

We can answer these questions by looking at the sampling distributions for s^2 and s, found by imagining that we compute them for many randomly generated data sets.

Sampling Distributions of s^2 and s

Histograms of s^2 and s computed from 10000 samples of independent, normal data points with $\mu = 0$ and $\sigma = 3$, for $n = 5$ and $n = 50$:



Are s^2 and s Unbiased Estimators?

The mean of the sampling distribution for s^2 turns out to be equal to σ^2. So s^2 is an *unbiased* estimator of σ^2.

This is why we divide by $n-1$ when computing s^2. If we divided by n, it wouldn't be unbiased.

However, s is *not* an unbiased estimator for σ. The mean of the sampling distribution for s is a bit smaller than σ. It's not far off, however, and the bias approaches zero as n gets bigger, so people don't bother to correct for this.

A Statistical Inference Problem

You are a "ham" radio operator who communicates with another operator in Mongolia. You try to use the signal delay to measure the distance, d, from your station to their station, using n measurements, $x_1, \ldots, x_n$.

From theory and past experience, you think the distribution of these measurements

- has mean equal to d.
- has a standard deviation of $\sigma = 100$ kilometres.

From $x_1, \ldots, x_n$, you compute $\bar{x} = (1/n) \sum_i x_i$.

What can you say about the distance d based on $\bar{x}$?

Sampling Distribution

Since the measurements are unbaised, we know that the mean of $\bar{x}$ is equal to d.

If the measurements are *independent*, the standard deviation of $\bar{x}$ will be $\sigma/\sqrt{n}$.

The mean and standard deviation tell us something about how accurate $\bar{x}$ is, but not everything.

The *sampling distribution* of $\bar{x}$ tells us more. It will be normal if the measurements are normally distributed. It will be approximately normal when n is large even if the distribution of the x_i is not normal.

Confidence Intervals

Using the sampling distribution, we can try to construct a C% *confidence interval* (C.I.) for d. A C.I. is a range (low, high) computed from $x_1, \ldots, x_n$ by a method that ensures that:

> If we compute the C.I. (low, high) many times, from many samples of size n, in the long run, C% of these intervals will contain d (ie, low $\leq d \leq$ high).

There are many different ways of computing confidence intervals that satisfy this, but when $\bar{x}$ has an approximately normal distribution, we usually use a confidence interval of the form $(\bar{x} - e,\ \bar{x} + e)$.

We need to set e so that this is indeed a C% confidence interval, for whatever *confidence level* C we choose.

Finding the Confidence Interval

Suppose that $\bar{x}$ is normally distributed with mean d and standard deviation $\sigma/\sqrt{n}$. Assume we know σ. How do we select e so that $(\bar{x} - e,\ \bar{x} + e)$ is a C% confidence interval?

We set e so that

$$P(\bar{x} > d + e) = P(\bar{x} < d - e) = (1 - C)/2$$

If this is so, then

$$P(\bar{x} - e \leq d \leq \bar{x} + e) = C$$

When the standard deviation of $\bar{x}$ is one, we can find such an e from the normal table. We just multiply to get the appropriate value for other standard deviations.

Note: We need to know σ, but we do *not* need to know the value of d. That's certainly fortunate!

Example Confidence Intervals

Here are the values of e to give a C.I. of $(\bar{x} - e,\ \bar{x} + e)$ for some commonly-used confidence levels:

90%:	$1.645\,\sigma/\sqrt{n}$
95%:	$1.960\,\sigma/\sqrt{n}$
99%:	$2.576\,\sigma/\sqrt{n}$

Suppose you decide to use a 95% confidence interval, and make $n = 16$ measurements, giving $\bar{x} = 5510$ kilometres. What is your confidence interval for the distance to the operator in Mongolia? (Recall that $\sigma = 100$.)

We find $e = 1.960 \times 100/\sqrt{16} = 49$. The 95% C.I. is $(\bar{x} - e,\ \bar{x} + e) = (5461,\ 5559)$.

What happens to the C.I. as we change C and n?