Sampling and inference

Frank Edwards

Large sample (asymptotic) theorems

As a sample of draws from a random variable increases, the sample mean converges to the population mean $E(X)$

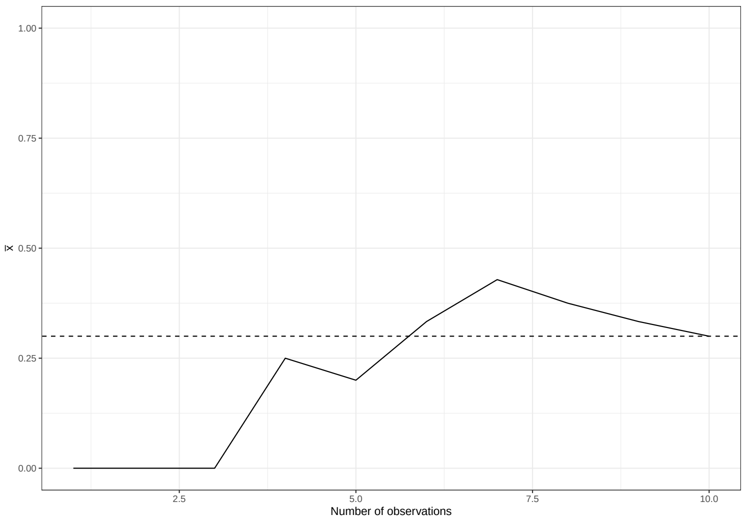
$$\bar{x}_n \to E(X)$$

Monte Carlo simulation for the mean of a binomial variable

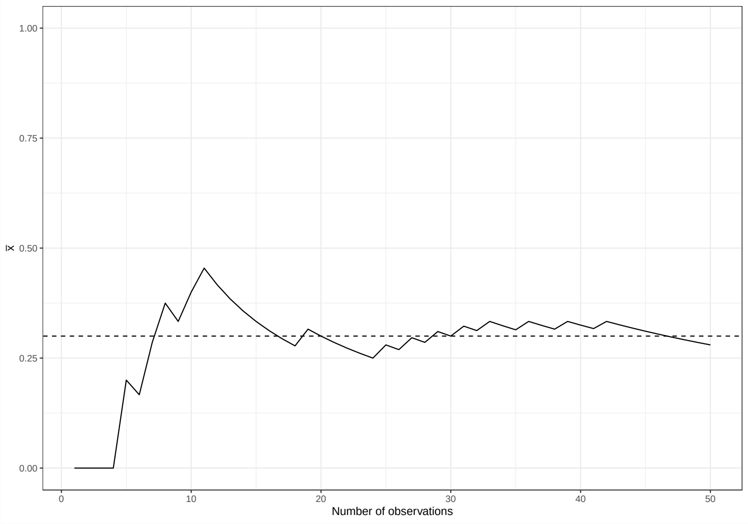
To test the law of large numbers, let's draw from $x \sim \mathrm{Bernoulli}(0.3)$ with varying sample sizes.

We expect that $\bar{x}$ will converge to $E(X)$ as the sample size n increases

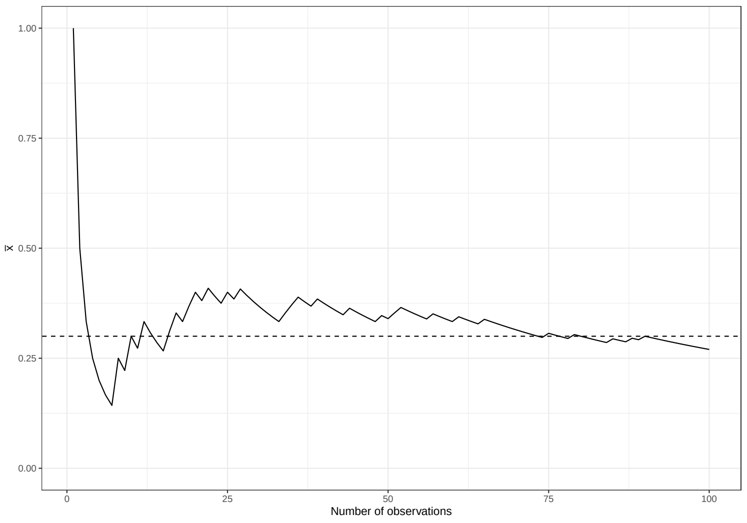
n = 1-10



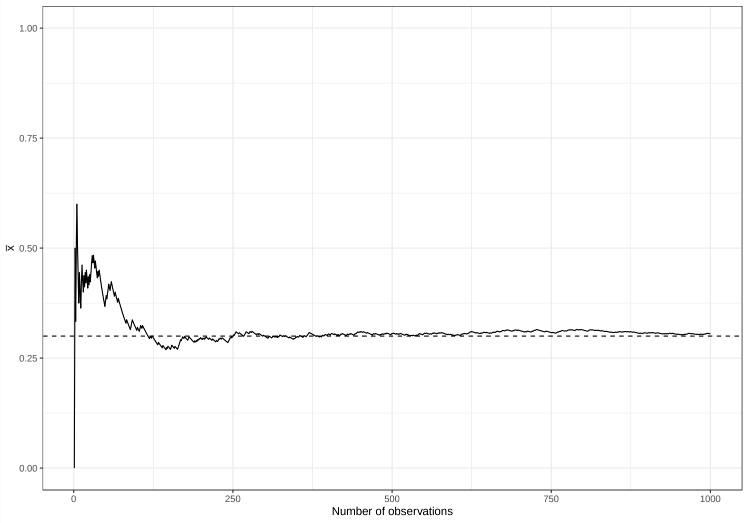
n = 50


1.00
0.75
0.50
0.25
0.00
$\bar{x}$
0
10
20
30
40
50
Number of observations

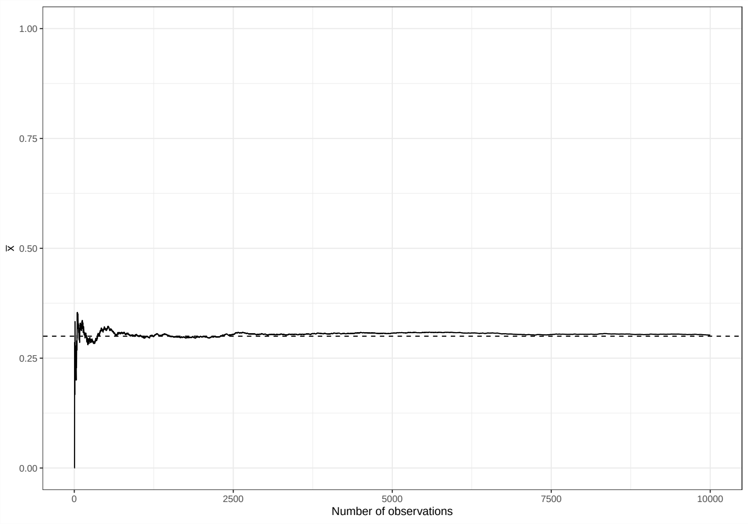
n = 100



n = 1000


1.00
0.75
0.50
0.25
0.00
$\bar{x}$
0
250
500
750
1000
Number of observations

n = 10000



1.00
0.75
0.50
0.25
0.00
$\bar{x}$
0
2500
5000
7500
10000
Number of observations

The Central Limit Theorem

If we assume samples $x_1, x_2, \cdots x_n$ are random samples from a population with mean μ, and $\bar{x_1}$ is the empirical mean of x_1 then:

- As n increases, the distribution of the sample means $\bar{x}$ approaches a Normal distribution with mean μ.

The Central Limit Theorem: implications

- This relationship holds for many distributions (Bernoulli, Binomial, Normal, others we'll discuss later)

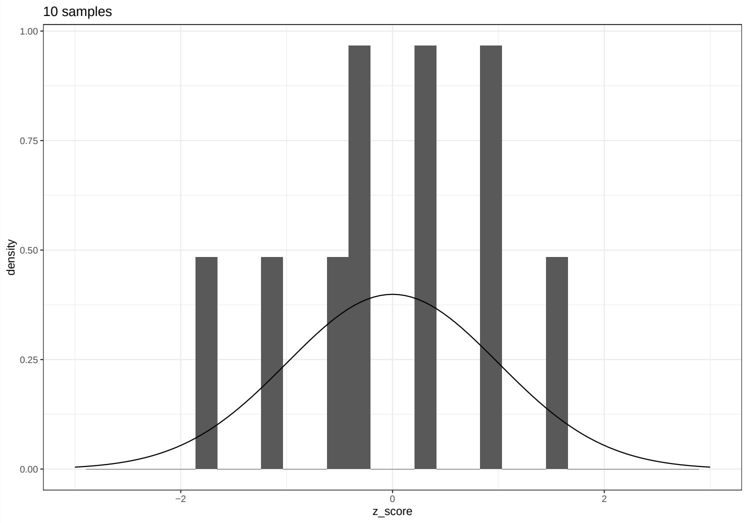
The Central Limit Theorem: implications

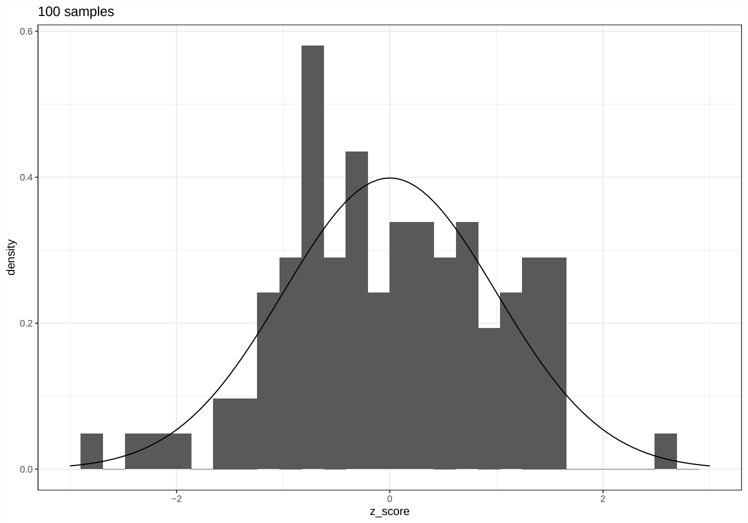
- This relationship holds for many distributions (Bernoulli, Binomial, Normal, others we'll discuss later)
- The distribution of z-scores of sample means converges to a $Normal(0, 1)$ distribution

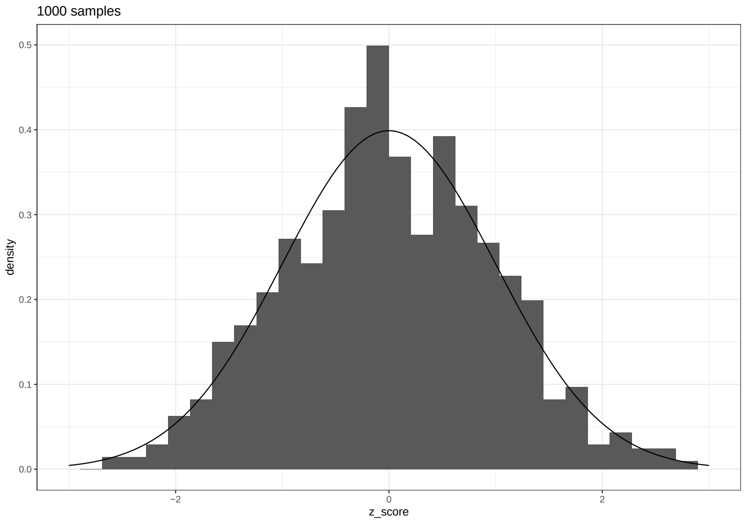
The Central Limit Theorem: implications

- This relationship holds for many distributions (Bernoulli, Binomial, Normal, others we'll discuss later)
- The distribution of z-scores of sample means converges to a $Normal(0, 1)$ distribution
- The Central Limit Theorem allows us to make statements about uncertainty when we haven't observed the population mean or variance

Monte Carlo simulations of a binomial variable p=0.7, n=10




100 samples
density
0.6
0.4
0.2
0.0
−2
0
2
z_score


1000 samples
0.5
0.4
0.3
0.2
0.1
0.0
density
−2
0
2
z_score

Exercise - simulation and the central limit theorem

- Let's write a quick sampler as a group to test the central limit theorem
- Our variable is $y \sim Binomial(10, 0.5)$
- So we need to 1) construct many random samples of y and compute the mean and 2) visualize the distribution of these means
- Assume the central limit theorem is true. Draw 100 samples of y and make an inference about the true parameter $\mu = np$ based on your observations of $\bar{y}$

Point estimates

$\bar{y}$ is our point estimate for the mean of the random variable y.

A point estimate is a *statistic* that reflects our best guess about the location of an unknown *parameter*

But point estimates alone are incomplete.

Uncertainty

Uncertainty statements quantify how much information we have about a statistical parameter. Uncertainty statements communicate information about the *precision* of our point estimate.

Return to our simulation exercise

- How much *uncertainty* do we have about the mean of our random variable when we simulate 10 trials?
- Visualize the distribution of $\bar{y}$ with 10 simulations. Now with 100 simulations. What do you notice?
- What summary statistic could help us quantify uncertainty?
- Compute this statistic for $\bar{y}$ with 10 simulations, and again with 100 simulations. What do you notice?

The standard deviation

Recall that we define a standard deviation as

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu^2)}{n}}$$

We will define σ as the **parameter**, and s as the estimated **statistic** for the standard deviation

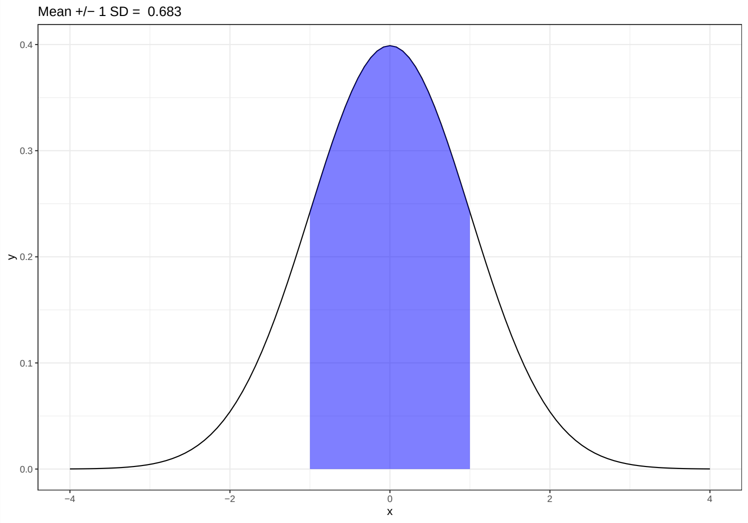
How certain are we about our estimate of the mean of the random variable?

- We know that for a sufficiently large sample $\bar{y} \sim \mathrm{Normal}(\mu, \sigma^2)$
- Which part of this claim can be attributed to the law of large numbers, and which part can be attributed to the central limit theorem?

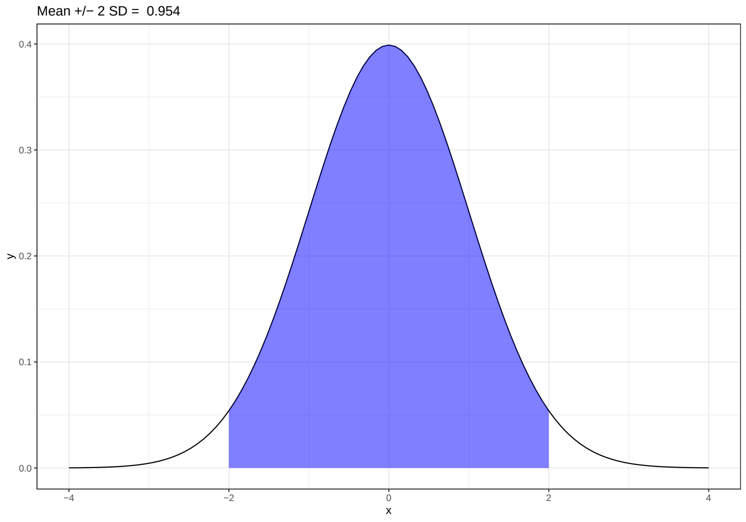
Confidence intervals: the basic logic

If we know that $\lim_{x \to \infty} \bar{y} = \mu$ (law of large numbers) and that the distribution of our point estimate $\bar{y}$ is approximately Normal for a large sample (central limit theorem), we can make *inferences* about the possible location of an unknown *parameter* using our estimated *statistics*

Return to the Normal PDF



Return to the Normal PDF



Defining our uncertainty bounds

95% is a conventional threshold to use for uncertainty (though there's nothing magic about it!)

To obtain 95% of a Normal pdf with a center at the mean (symmetric) we can simply compute

```
# compute location below which 2.5% of the Normal PDF falls
qnorm(0.025, 0, 1)
```

```
## [1] -1.959964
```

```
# compute location below which 97.5% of the Normal PDF falls
qnorm(0.975, 0, 1)
```

```
## [1] 1.959964
```

These two points define a symmetric region between which 95% of the area of the Normal(0,1) PDF lies.

Now, compute a confidence interval for our simulation

Let's begin with our simulation with 10 samples (each of which has 10 trials)

Recipe to bake a confidence interval

1. Compute the sample mean ($\overline{y}$)
2. Compute the standard deviation of the sample mean
3. Define your critical values (0.95 is conventional, resulting in critical values of +/- 1.96)
4. Compute $\overline{y} \pm 1.96 \times s$

What do you obtain?

What were our estimated intervals for 10 samples?

- No seriously, let's list them.
- Then plot them
- If we had enough of us in the class, 95% of these intervals would contain μ!

Interpretation of a confidence interval

- If we repeated the experiment many times, and computed many confidence intervals, 95% of the intervals would contain the **parameter** μ.
- There is NOT a 95% chance your interval contains μ. This is a subtle point that is often mistaken

Precision and confidence intervals

- Provide your confidence interval for μ with our simulation under n = 10
- Now n = 100
- Now n = 1000
- What is happening and why is it happening?