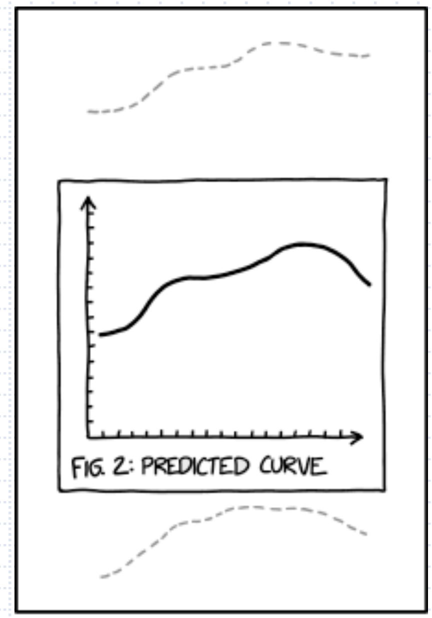
Distributions & the Central Limit Theorem 3



SCIENCE TIP: IF YOUR MODEL IS BAD ENOUGH, THE CONFIDENCE INTERVALS WILL FALL OUTSIDE THE PRINTABLE AREA.

The Central Limit Theorem Definition

- If you have a random variable that comes from a distribution with mean=μ and standard deviation=σ, the following is true for the sampling distribution (i.e. the distribution of all possible sample means) from samples of size n
  - If n is large enough, the shape of the sampling distribution will be approximately normal
  - The mean of the sampling distribution is μ
  - The standard deviation of the sampling distribution is $\sigma/\sqrt{n}$

Central Limit Theorem:
Why do we care?

- If we know that the distribution of $\bar{X}$ is normal, then we can use that to calculate the probability of obtaining an $\bar{X}$ as extreme as the one we did (i.e. that we got $\bar{x}$ as our sample mean)

- For example, we can find things like P($\bar{X} \geq \bar{x}$) because $\bar{X}$ follows a normal distribution

- We can compute the standard normal (z-scores) from this.

Sampling distributions

- Does this make sense?
  - It makes sense that the distribution of means would cluster around the population mean
  - It makes sense that the variability in the means is smaller than in the raw data because the extreme values are already averaged out (remember $\sigma/\sqrt{n}$)
  - The part about the distribution being normal if n is large enough? Mathematical proof … But we can demonstrate it for several examples

Sampling distributions
Distribution of the Sample Mean

- σ is the standard deviation of the original distribution
- $\sigma/\sqrt{n}$ is called the *standard error,* or more precisely,

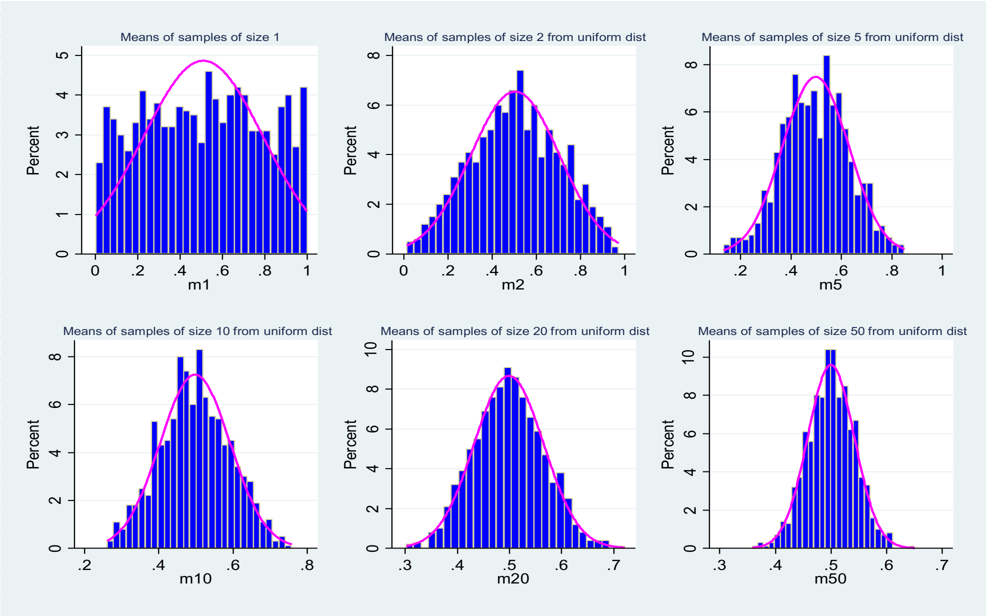
the *standard error of the mean*,

and it is the standard deviation of

the distribution of the sample mean

The Uniform Distribution where

each value has the same probability



Uniform Distribution Example

- Taking samples from the following Uniform distribution
- Uniform(1, 100) with different sample sizes
- Try this with one of these applets to convince yourself that the distribution of the sample mean approaches a Normal distribution

- http://www.ltcconline.net/greenl/java/Statistics/clt/cltsimulation.html

- http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html

Using the Central Limit Theorem (CLT)

- Suppose we sampled from an HIV infected population with a known mean CD4 count, μ, of 250 cells/mm^3 and a standard deviation, σ, of 200 cells/mm^3
- If we select repeated samples of size 50, what proportion of the samples will have a mean value less than 100 cells/mm^3?
- Using the CLT, we know the mean of each of the samples, $\bar{x}$, is a random variable that follows a Normal distribution with a mean of 250 and a standard error of $\sigma/\sqrt{n} = 200/\sqrt{50}$
- In STATA di 200/sqrt(50)

  28.284270

Using the Central Limit Theorem

- So $\bar{X}$ is ~ Normal(250,28.3)
- Then we know that ($\bar{x}$ - 250)/28.3) ~ N(0,1)
- To find the proportion of means <100, we calculate

$$(100 - 250)/28.3 = -150/28.3 = -5.3$$

So what is the probability P(Z<-5.3)?

in STATA:

di normal(-5.3)

5.790e-08   very very small probability

Using the Central Limit Theorem

- What level of CD4 count is the <u>lower</u> 2.5th percentile of the mean values?
- For what value of z is P(Z<z)=.025 ?

```
di invnormal(.025)
-1.959964
```

- Now we use this to get back to $\bar{X}$, using $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$
- So -1.96=($\bar{x}$-250)/(200/$\sqrt{50}$)
- Rearranging, $\bar{x}$= [(-1.96*200)/$\sqrt{50}$]+250

```
di -1.96*200/sqrt(50)+250
194.56283
```

Using the Central Limit Theorem

- What level of CD4 count is the upper 2.5^{th} percentile of the mean values?
- What is the z-value that solves P(Z>z)=0.025?
- Remember invnormal gives use the z-value for P(Z<z)=p, so we need to use 1-p in the invnormal command

```
di invnormal(1-.025)
1.959964
```

- So $1.96=(\bar{x}-250)/(200)/\sqrt{50})$
- Rearrange to solve for $\bar{x}$

```
di 1.96*200/sqrt(50)+250
305.43717
```

Using the Central Limit Theorem

- Now we have the lower and upper 2.5% percentiles <u>of the distribution of the sample means</u>.
- The interior area contains 95% of the sample means.
- 95% of the means from samples of size 50 that come from the underlying distribution ~N(250,200) will lie within this interval (194.6, 305.4)

This is the 95% confidence interval around the mean

Using the Central Limit Theorem

- If we selected just one sample of size 50 (what you usually do) and the sample mean was outside these limits (e.g. 315), it possibly came from a different population, or that a rare (<5% probability) event had occurred.
  - Because we had said that 95% of the time the sample mean will be in the range of 195-305
  - We could say this because the central limit theorem told us that the distribution of the sample means is approximately a normal distribution with mean 250 and standard deviation 200/ *√50*
  - This would change if our sample size were different (i.e.; 300 instead of 50)
    - di 1.96*200/sqrt(300) and di -1.96*200/sqrt(300)
    - 227.37 to 272.63

As the sample size increases, the confidence interval narrows!

Recap on the Central Limit Theorem

- Assures us that taking one sample of size n provides the ability to make inferences about our data
- The sample mean is the expected value of the population mean
- The standard deviation of the sample mean, called the standard error, provides an estimate of the variability of the sample mean
- We can use STATA to calculate the probability or the values of a confidence interval
- As the sample size increases, the width of the interval decreases