Confidence Intervals of the mean

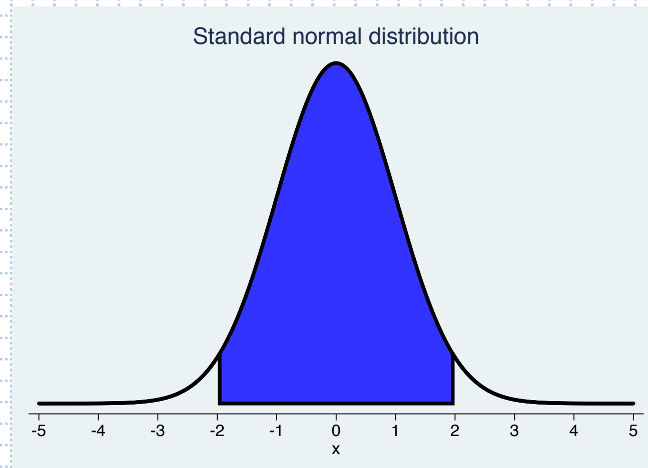
Confidence intervals for means

- The sample mean, $\overline{X}$, is the point estimate of μ, the population mean
- Different samples will yield different $\overline{X}$'s so we cannot be sure how our estimate differs from μ
- Interval estimation provides a range of reasonable values that contain the population value with a certain degree of confidence
- This is the definition of the confidence interval combining what we know from the Normal distribution and the Central Limit Theorem
- By the CLT, $\overline{X}$ follows a normal distribution and if n is sufficiently large, z follows a standard normal distribution

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Confidence intervals for means

- We use what we know from the CLT to calculate confidence intervals for means in general
- We know from examining the standard normal distribution that $P(-1.96 \leq Z \leq 1.96) = 0.95$
- Substituting the formula for Z into the above we get $$P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96) = 0.95$$



- Rearranging and multiplying by -1 within the parentheses we get:

$$P(\bar{X} - 1.96\sigma / \sqrt{n} \leq \mu \leq \bar{X} + 1.96\sigma / \sqrt{n}) = 0.95$$

Confidence intervals for means

- Thus the lower 95% confidence limit for μ is $\bar{X} - 1.96\sigma / \sqrt{n}$
- And the upper 95% confidence limit for μ is $\bar{X} + 1.96\sigma / \sqrt{n}$
- We say we are 95% confident that the interval we calculate using the above formulae includes $\boldsymbol{\mu}$ (the true mean of the distribution)

- $\bar{X}$ is a random variable that is the result of sampling and μ is a population parameter that is fixed in perpetuity; it as the same value irrespective of the sample
- μ is either in the interval you calculate or it is not

<u>What is random is the interval</u> because it is based on the sample

Confidence intervals for means

- So we say that the probability that the interval contains the true population mean is 95%
- If we were to select 100 random samples from the population and calculate confidence intervals for each, approximately 95 of them would include the true population mean μ (and 5 would not)

90% confidence interval

Replace 1.96 in the formula with 1.64

99% confidence interval

Replace 1.96 in the interval with 2.58

Confidence intervals for means

- How to get a tighter interval? Decrease the confidence level.
- How to get a tighter interval? Increase the sample size.

Confidence level	α	$Z_{\alpha/2}$
99%	.01	2.58
95%	.05	1.96
90%	.10	1.64
80%	.20	1.28

n	95% confidence limits	Length of interval
10	$\bar{x} \pm 1.96\sigma/\sqrt{10} = \bar{x} \pm 0.620\sigma$	1.240σ
100	$\bar{x} \pm 1.96\sigma/\sqrt{100} = \bar{x} \pm 0.3920\sigma$	0.784σ
1000	$\bar{x} \pm 1.96\sigma/\sqrt{1000} = \bar{x} \pm 0.062\sigma$	0.124σ

Confidence intervals for means

- What to do when σ is not known? (In practice, always)
- By the Central Limit Theorem, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ follows a normal distribution, if n is sufficiently large
- Can we substitute s, the sample standard deviation for σ ? Sometimes
- **s** is not a reliable estimate of **σ** if n is small

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

- If X is normally distributed, and a sample of size n is chosen, then

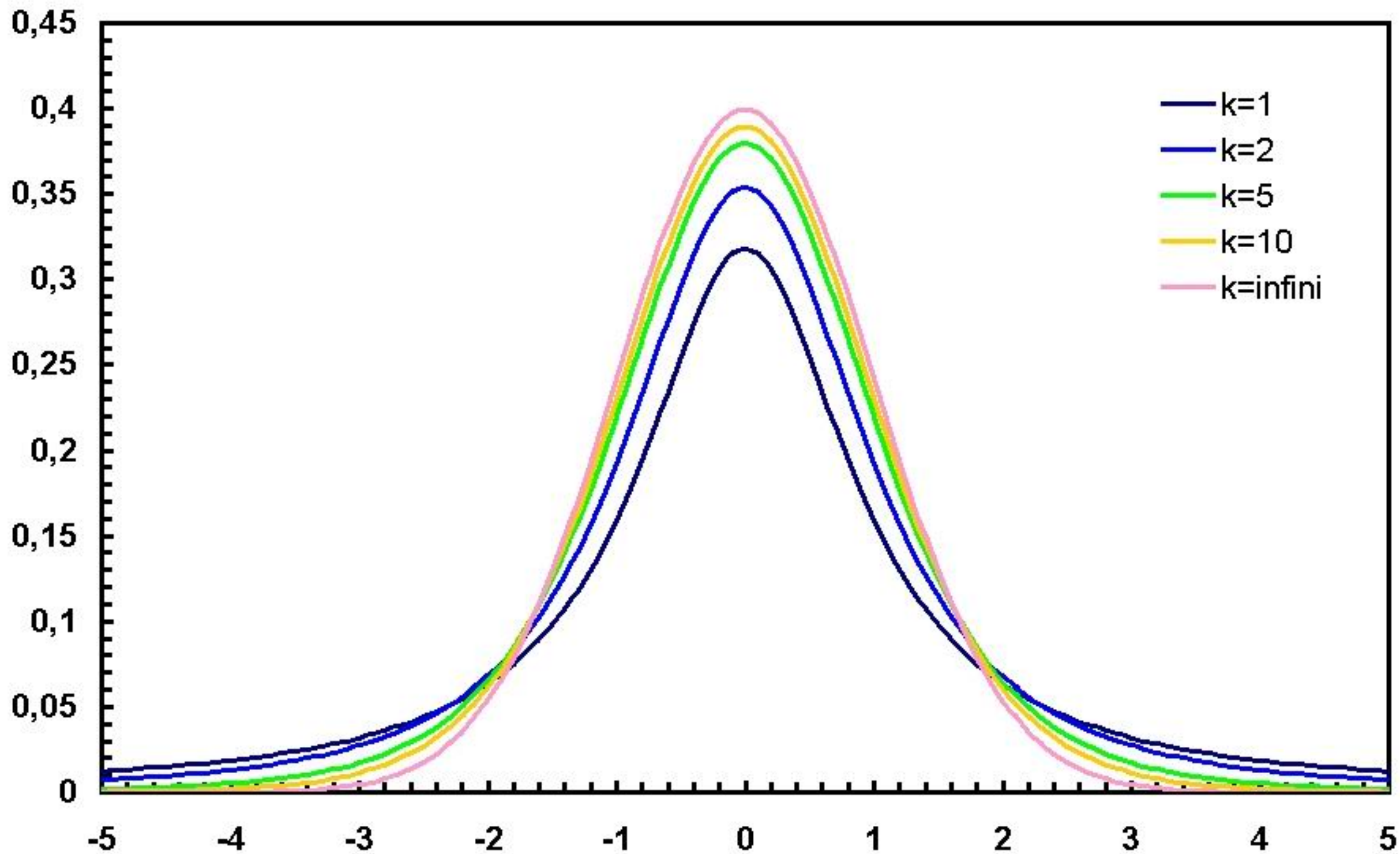
$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$ follows a Student's t distribution with n-1 degrees of freedom

Student's t distribution: Who was Student*?

- The mean of the t distribution is 0 and the standard deviation is 1
- The t distribution is symmetric and bell-shaped, but has heavier tails than the standard normal – extreme values are more likely to occur
- For small n, the tails are fatter
- For large n, the t distribution approaches (i.e. becomes indistinguishable from) the standard normal distribution

*Will Gosset

The t-distribution

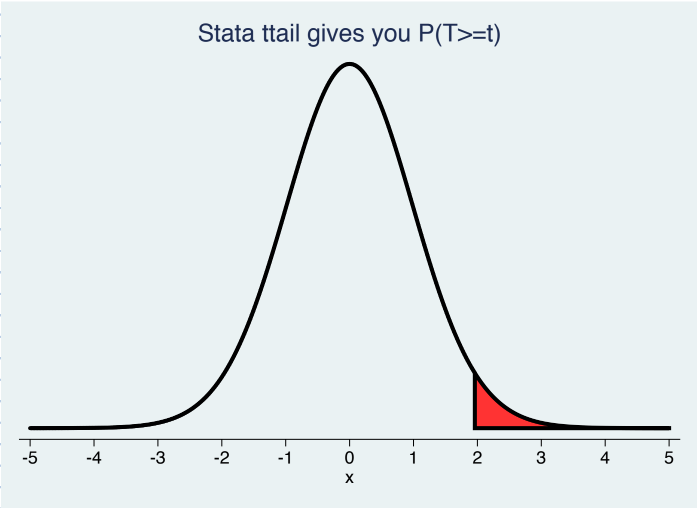


Student's t-distribution

- There are separate curves for each degree of freedom (df)
- In STATA: P(T≥t) is calculated using ttail *****
- The code is ttail(df,t)
- e.g., P(T>1.96) n=20 display ttail(19,1.96) .03241449

<u>USE n-1 for the df</u>

<u>Recall: What are degrees of freedom?</u>



Student's t-distribution

- To find the value for which P(T>t)=p use invttail(df,p) in Stata
- For example, for what t is P(T>t)=.025 for a sample of size 20?
- The answer for this t cutoff value is denoted $t_{19,.05}$

  display invttail(19,.025) 2.0930241

- 95% confidence intervals for means when σ is not known
- So using the t-distribution, the formula for a for a 95% confidence interval for a mean is:

$$(\bar{X} - t_{df,0.025} s / \sqrt{n}, \bar{X} + t_{df,0.025} s / \sqrt{n})$$

where df =n-1

Confidence intervals for means

- Remember that when n is large, the t distribution approaches the normal distribution
  - E.g. $z_{0.025} = 1.960$
  - While $t_{n-1,0.025} =$ →

n	$t_{n-1,0.025}$
2	**12.706**
3	**4.303**
5	**2.776**
10	**2.262**
50	**2.010**
100	**1.984**
200	**1.972**
300	**1.968**
500	**1.965**
1000	**1.962**
1500	**1.962**

Confidence intervals for means: Example

- CD4 cell count among HIV positives diagnosed at Mulago Hospital
  - N=999
  - Sample mean = 329.2
  - Sample SD = 266.1
  - t cutoff?

```
. di invttail(998,.025)
1.9623438
```

  - 95% CI

$$= (329.2 - 1.962*266.1/\sqrt{999},\ 329.2 + 1.962*266.1/\sqrt{999})$$

$$= (312.7\text{-}345.7)$$

Confidence intervals for means in STATA

```
. . summ cd4count

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    cd4count |       999    329.2332    266.1177          1       1932


. mean cd4count

Mean estimation                     Number of obs    =     999

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    cd4count |   329.2332   8.419592      312.7111    345.7554
--------------------------------------------------------------


. ci mean cd4count

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    cd4count |        999    329.2332    8.419592        312.7111    345.7554
```

Confidence intervals for means in STATA

- Note that some statistical output gives you the SE or the SEM, which stands for standard error or standard error of the mean.
- This is $s/\sqrt{n}$ which is what you will use in confidence intervals
  - *So what if we have Binomial data?*