

3190 Week 5

Normality & confidence intervals

Source of confusion

- We can use Shapiro Wilks to test for normality
- But we keep hearing that with representative samples of $n \geq 30$, normality can be assumed
- So why do normality tests?
- Remind me to answer this later...

Basic terminology

- We have only used sample symbols to date

\bar{X} = sample mean

μ = population mean

S^2 = sample variance

σ^2 = population variance

S = sample standard deviation

σ = population standard deviation

n = sample size

N = population size

- You need to be able to recognize that these symbols mean close to the same thing, fluidly
- You must also be clear on the difference between them

To assume normality

- We are not concerned about the shape of the sample and its scores, just that the sample is random **so that error is random**
- We are concerned about the **error distribution**; aka called the sampling distribution of the statistic (can be mean, s, etc...)
- Another layer of quantification

The distribution of error

- 5 layers of quantification here
 - 1) raw data scores
 - 2) z-scores calculated from raw scores
 - 3) area under curve related to z-score
 - 4) area equals *probability of encounter* in a distribution
 - 5) error distribution of the statistic

The Sampling Distribution of the Sample Mean

- The distribution of **all possible** sample means from a population for a variable at a particular sample size
- Each sample produces an x-bar; each x-bar is an estimation of μ
 - Estimates are not quite μ
 - This is called **error**
- The sampling distribution of the sample means is an error distribution

Problem

- This is tough to communicate because we have to examine thousands of sample means to show it
- Instead, let's use a very small population and small samples to illustrate the point
- Heights of 5 starting basketball players = the population; we will sample it

An example

Player 1	Player 2	Player 3	Player 4	Player 5
76 inches	78 inches	79 inches	81 inches	86 inches

What is μ ? It = 80 inches

Let's take a sample of 2 randomly

We choose players 2 & 5. What is \bar{x} ? It = 82 inches

Can you see that \bar{x} is an estimate of μ ?

If we take all possible samples of 2 and determine \bar{x} for each one; we have the sampling distribution of the sample means.

What would happen if our sample was larger?

Example, cont'd

Sample	Heights	x-bar
1,2	76, 78	77.0
1,3	76, 79	77.5
1,4	76, 81	78.5
1,5	76, 86	81
2,3	78, 79	78.5
2,4	78, 81	79.5
2,5	78, 86	82.0
3,4	79, 81	80.0
3,5	79, 86	82.5
4,5	81, 86	83.5

The sampling distribution of the sample means

It is all possible sample means for sample size of 2

These are all estimates of μ ; only one of the x-bars is an exact estimate (sample 3,4)

We are very interested in the distribution of these estimates because all inferential tests compare estimates (often of μ)

New thinking

- We will no longer be thinking in terms of samples and their distributions
- We will shift to the *error distributions of estimates*; for example the sampling distribution of the sample means
 - Statistics are estimates
- **Why?** Because we are interested in comparing statistics from different samples as estimates to see how likely they are from the same error distribution

Standard Error

- This is the **standard deviation of the sampling distribution of sample means**
 - S_E = standard deviation of the sampling distribution of the sample means
- It answers “how many standard deviations away from μ is an particular \bar{x} ?”
- How many standard deviations is \bar{x} for sample 2,4 from μ ?

Calculating S_E

- Since we cannot draw thousands of sample from a population to get S for the distribution of \bar{x} ...
 - We must estimate S_E
 - It turns or this is quite easy to do

$$S_E = \frac{S}{\sqrt{n}}$$

← both are standard error →

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

σ = population standard deviation

- We know this from experiments with populations and samples

Standard Error

- Often depicted as the standard deviation of \bar{x}
- σ = the population standard deviation
- When σ is known, use it
- It is often unknown so we estimate using S

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Standard Error

- Let's inspect this equation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- What makes SE smaller?

$$S_E = \frac{s}{\sqrt{n}}$$

- Bigger sample size
- So, error (S_E) is smaller with larger sample size

The central limit theorem

- Sufficiently large random samples for a variable produce a normal-shaped error distribution
- At $n \geq 30$, in random samples, the error distribution for \bar{x} is normal, matter what the shape of the population
- We can then use S_E to assess confidence in our \bar{x}
- This is why $n \geq 30$ is the magic number

So what do you need to know...

- That we do not have to base parametric statistics on data distributions directly...
 - But on underlying **error distributions**
- That if samples are random, at $n \geq 30$ normality can be assumed because error distributions are normal at that size
- That in parametric tests we compare error distributions, such as the sampling distribution of the sample means

Confused?

- You can remain confused about the underlying math if you are willing to accept it...
- What does it mean to be confused, incidentally?
- What's the root of the word...?

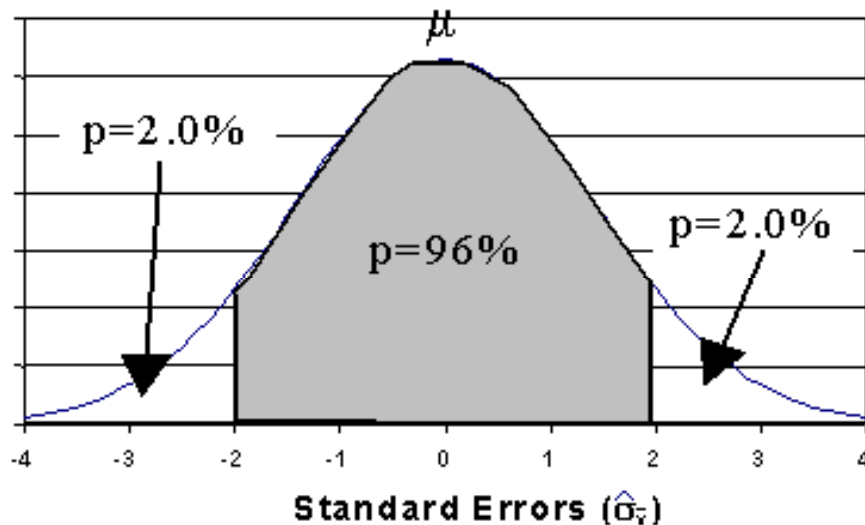
The value of confusion...

- Confusion can be useful
- It can be a warning sign...
- It can signal that you are close to learning something new...
- It can be a sign that it is time to ask a question
- It can be a good thing, if you use it wisely...

- Embrace it and take the plunge...

From Confusion to Confidence!

- If you know σ or S , you can provide confidence intervals using the standard error & the z distribution



At μ $z = 0$

An \bar{X} above μ has a $+\sigma_{\bar{x}}$

An \bar{X} below μ has a $-\sigma_{\bar{x}}$

Confidence intervals

- I take one sample randomly from a population
- I want to know with 90% confidence that the sample mean, \bar{X} falls within a certain range of μ
- I can use the normal distribution to figure this out
- Go to the normal table ± 1.65 Z contains about 90% of the area
- Multiply $\pm 1.65 \times \sigma_{\bar{x}}$ to get the 90% confidence interval
- What about 95% confidence intervals? What z-score do you use?

Example

- Average journey to work is 9.6 miles in 50 commuters ($n = 50$)
- Population standard deviation *thought from other studies* to be about 3 miles

$$\bar{X} \pm Z \sigma_{\bar{x}} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 0.42$$

- $\bar{X} = 9.6$; $\sigma = 3$
- What is the 90% CI? $9.6 \pm 1.65(0.42) = 9.6 \pm 0.69$
- So, you are 90% certain that μ is between 8.91 and 10.29 miles for a commute
 - Because that is 9.6 ± 0.69

What did we just do?

- We used the normal curve to offer probability of 90% using $z = 1.65$
- We used the error distribution for \bar{X} ($\sigma_{\bar{x}}$) at $n = 50$ to place our \bar{X} (9.6 miles) in reference to μ ($z = 0$)
- We don't know μ , but we know σ and n , *and we know the z-score for $\mu = 0$*
- We scaled our $\sigma_{\bar{x}}$ to $z = 1.65$ and added/subtracted it from \bar{X} to provide the 90% CI

Confidence intervals

- If \bar{x} estimates μ , then how much error is in the estimate given sample size?
- Put \bar{x} at $z = 0$ and determine, given sample size the probability μ falls within a certain range of it
 - S_E is a product of sample variability (S) and sample size sqrt of n

$$S_E = \frac{S}{\sqrt{n}}$$

To Review

- We have shifted from describing sample scores with S in reference to \bar{X}
- to using the same scale (z-scores) to describe \bar{X}_s with $\sigma_{\bar{X}}$ in reference to μ
- It is the *same thing*;
 - the first is a data distribution with case scores,
 - the second is an error distribution with all possible \bar{X}_s

What do you need to know?

- To calculate CI when σ is known
 - How to use the normal table
 - How to calculate $\sigma_{\bar{x}}$
 - How to get the z-score for a CI
 - How to conceive of $\sigma_{\bar{x}}$ related to μ on the normal curve
 - It helps if you understand what the *sampling distribution of the sample means* is

When you do not know σ ...

- When we do not know σ we must *estimate* $\sigma_{\bar{x}}$ using S
- This introduces *error* because we are using a *sample characteristic in place of the population's*
- We will symbolize the standard error calculated with S as S_E

$$S_E = \frac{S}{\sqrt{n}}$$

- It turns out that we usually have to do this

The t distribution

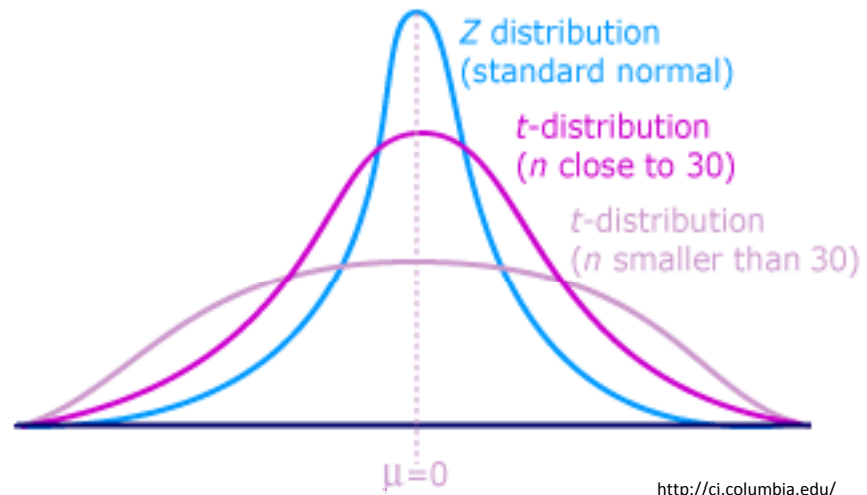
- Because we introduce error when we use S instead of σ , we cannot use the normal distribution to determine CI
- We must use a different distribution, **the t-distribution**
- It assumes a different shape at smaller sample sizes

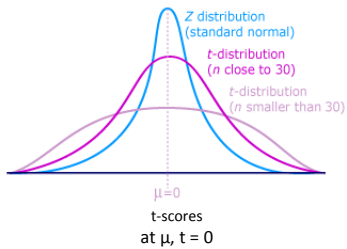
t-distribution

- You can see how the shape of the t distribution flattens at smaller sample size
- Why? Because the S_E is an estimate based on a sample

$$S_E = \frac{S}{\sqrt{n}}$$

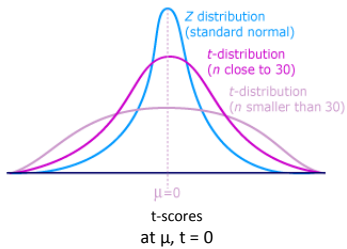
- Error in S is greater at small sample sizes, so the curve widens (makes it hard to be near the middle, μ)





What happens with t...

- Because the t-distribution is wider at smaller sample sizes
 - The probability of encountering an \bar{x} near μ **decreases**
 - That is, in a z distribution most of the \bar{x} are near μ and $\sigma_{\bar{x}}$ is small
 - At small sample sizes, S_E gets **larger**
 - Thus, it is harder to provide tight confidence intervals at smaller sample sizes
 - *We want this because it is conservative*

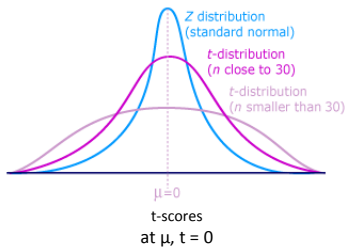


Degrees of Freedom

- Sample size (sort of...)

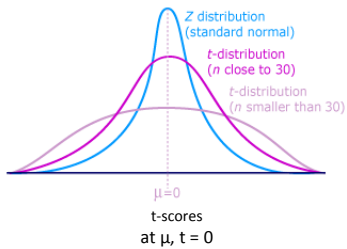
- “At the moment, I'm inclined to define **degrees of freedom** as a **way of keeping score**. A data set contains a number of observations, say, n . They constitute n individual pieces of information. These pieces of information can be used either to estimate parameters or variability. In general, each item being estimated costs one degree of freedom. The remaining degrees of freedom are used to estimate variability. All we have to do is count properly.” <http://www.jerrydallal.com/LHSP/dof.htm>

- “In short, think of df as a mathematical restriction that we need to put in place when we *calculate an estimate* one statistic from an *estimate* of another.” http://www.statsdirect.com/help/basics/degrees_of_freedom.htm



The t-table

- In the z table there is one area under the curve for each z score
- In the t-table the area under the curve varies with df (sample size)
- S has more error at lower df
- At lower df there is less area near $t = 0$ because the curve is flatter
- This compensates for higher error using S to estimate S_E by making it tougher to confidently predict μ



Confidence Intervals with t

- When we estimate S_E using S we use t to provide confidence intervals because of the error in S

$$\bar{X} \pm t(S_E) \quad \text{when} \quad S_E = \frac{S}{\sqrt{n}}$$

An example

- Average journey to work is 9.6 miles in 10 commuters ($n = 10$)
- σ is unknown, so use S for standard error

- $\bar{X} = 9.6; s = 2.5$
- What is the 90% CI?

$$\bar{X} \pm t(S_E) \quad \text{when} \quad S_E = \frac{S}{\sqrt{n}}$$

- So, you are 90% certain that μ is between ? and ? miles for a commute

Using the t-table

TABLE C

Student's t Distribution

t	Degrees of freedom									
	1	2	3	4	5	6	7	8	9	
0.1	0317	0353	0367	0374	0379	0382	0384	0386	0387	
0.2	0628	0700	0729	0744	0753	0760	0764	0768	0770	
0.3	0928	1038	1081	1104	1119	1129	1136	1141	1145	
0.4	1211	1361	1420	1452	1472	1485	1495	1502	1508	
0.5	1476	1667	1743	1783	1809	1826	1838	1847	1855	
0.6	1720	1953	2046	2096	2127	2148	2163	2174	2183	
0.7	1944	2218	2328	2387	2424	2449	2467	2481	2492	
0.8	2148	2462	2589	2657	2700	2729	2750	2766	2778	
0.9	2333	2684	2828	2905	2953	2986	3010	3028	3042	
1.0	2500	2887	3045	3130	3184	3220	3247	3267	3283	
1.1	2651	3070	3242	3335	3393	3433	3461	3483	3501	
1.2	2789	3235	3419	3518	3581	3623	3654	3678	3696	
1.3	2913	3384	3578	3683	3748	3793	3826	3851	3870	
1.4	3026	3518	3720	3829	3898	3945	3979	4005	4025	
1.5	3128	3638	3847	3960	4030	4079	4114	4140	4161	
1.6	3222	3746	3960	4075	4148	4196	4232	4259	4280	
1.7	3307	3844	4062	4178	4251	4300	4335	4362	4383	
1.8	3386	3932	4152	4269	4341	4390	4426	4452	4473	
1.9	3458	4026	4232	4349	4421	4469	4504	4530	4551	
2.0	3524	4082	4303	4419	4490	4538	4572	4597	4617	
2.1	3585	4147	4367	4482	4551	4598	4631	4655	4674	
2.2	3642	4206	4424	4537	4605	4649	4681	4705	4723	
2.3	3695	4259	4475	4585	4651	4694	4725	4748	4765	
2.4	3743	4308	4521	4628	4692	4734	4763	4784	4801	
2.5	3789	4352	4561	4666	4728	4767	4795	4815	4831	

For area of 0.45 at df = 9, t is between 1.8 and 1.9

$$t = 1.85$$

$$CI = \bar{X} \pm t(S_E) \quad S_E = \frac{S}{\sqrt{n}}$$

$$9.6 \pm 1.85(0.79) = 9.6 \pm 1.46$$

$$90\% CI = 8.14 \text{ to } 11.06$$

Compare to the previous example

- 9.6 ± 0.69 at $n = 50$, $\sigma = 3.0$; CI = 8.91 to 10.29
- Here, 9.6 ± 1.46 at $n = 10$, $S = 2.5$; CI = 8.14 to 11.06
- Why are we less confident in the second example?
 - 1) we estimated σ from S to get S_E
 - 2) we had smaller sample size ($n = 10$)
- As a result we had to use the t distribution, with a wider curve (equals broader area around μ)

What you need to know...

- When to use the t-table
 - = When we use S to estimate S_E
- How to use the t-table
 - What are degrees of freedom
 - Values change with n (df)
- How to express confidence that $\bar{X} = \mu$
- The logic behind using t instead of z .

Confidence Intervals in SPSS

The screenshot shows the SPSS Data Editor interface with the 'Explore' menu open. The 'Explore...' option is highlighted, and a red arrow points to the 'Statistics...' button in the 'Explore' dialog box. The 'Explore: Statistics' dialog box is also shown, with the 'Confidence Interval for Mean' set to 95%. A blue arrow points from the 'OK' button in the 'Explore: Statistics' dialog to the text 'Push OK once you have entered the CI level you desire'.

Sex	Age
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5
B	.5

Explore

Dependent List: Weight

Factor List:

Label Cases by:

Display: Both Statistics Plots

Statistics... Plots... Options...

Explore: Statistics

Descriptives

Confidence Interval for Mean: 95 %

M-estimators

Outliers

Percentiles

Continue Cancel Help

OK Paste Reset Cancel Help

Push OK once you have entered the CI level you desire

Confidence Intervals in SPSS: Output

Descriptives

		Statistic	Std. Error
Weight	Mean	60.3521	.34842
	95% Confidence Interval for Mean	Lower Bound	59.6689
		Upper Bound	61.0353
	5% Trimmed Mean	59.7616	
	Median	61.0000	
	Variance	329.226	
	Std. Deviation	18.14458	
	Minimum	17.00	
	Maximum	131.00	
	Range	114.00	
	Interquartile Range	21.00	
	Skewness	.326	.047
	Kurtosis	.320	.094