<<Lecture6_Student.pdf>>

Stat 324 – Introduction to Statistics for Engineers

LECTURE 6: ESTIMATING μ: THE POPULATION **MEAN** WITH CONFIDENCE

SECTION 5.2, 5.3 OF OTT AND LONGNECKER

Random Variables in Sampling and Estimation

Motivating Example:
Ex 1: The manufacturer **wants to know the average amount of paint applied by the device**, so 16 blocks are selected at random, and the paint thickness is measured in mm. The results are below:

1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83, 1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49

We can use the sample mean $\hat{\mu} = \bar{X} = 1.348$ as a point estimate for μ: the population **mean** paint thickness of all blocks

This single best guess is almost always wrong. ☹

But if the standard error of the estimate $SD(\bar{X})$ is small, then the estimate will usually be close

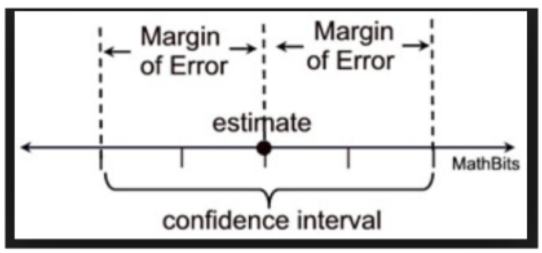
We will construct a **confidence interval (L,U) around** $\hat{\mu}$ to get a range of good guesses for μ:

How we will construct the confidence interval depends on the assumptions we want to make about the information we "know" about the population and the samples we observe.

Constructing Confidence Intervals for an Unknown μ and σ known

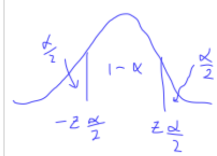
We would like to construct a confidence interval (L, U) where

- L and U are RVs and each confidence interval we create is just one realization
- $100\%(1-\alpha)$ of the CI realizations contain the true μ, where $\alpha \in (0,1)$
  - $100\%\,(1-\alpha)$ is the confidence level (90%, 95% most common)
  - $P(L \le \mu \le U) = 1-\alpha$ for a specified α (10% and 5% most common)
  - L and U are functions of the data we are able to collect from samples or know about the population.



"Margin of Error"
=
" half width **"**

Constructing Confidence Intervals for an Unknown μ and σ known



If we can make the assumption that $X_i \sim N(\mu, \sigma^2)$, then we know $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

(Or if n is large enough that the CLT kicks in, this is at least approximately true.)

Using critical values: $P\left(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}\right) = 1-\alpha$, (Draw)

Substitute $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$, $\quad P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha \quad$ So $P\left(\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

"$100\%(1-\alpha)$ of possible $\bar{X}$ are within $z_{\frac{\alpha}{2}}$ SE($\bar{X}$) or $z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ of μ"

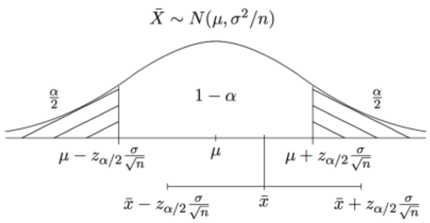
Rewrite: $P\left(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

Before we sample, the random interval from $L = \bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ to $U = \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ will contain the true μ $100(1-\alpha)\%$ of the time.

Constructing Confidence Intervals for an Unknown μ and σ known

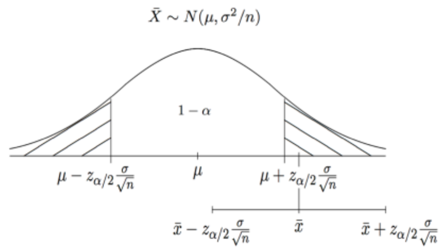
For this realization of $\bar{X} = \bar{x}$, μ is between l and u and therefore is "contained" in the CI.

This will happen with $100\%(1-\alpha)$ of the possible $\bar{X} = \bar{x}$ and therefore samples we could take



For this realization of $\bar{X} = \bar{x}$, μ is **NOT** between l and u and therefore is "not contained" in the CI.

This will happen with $100\%\alpha$ of the possible $\bar{X} = \bar{x}$ and therefore samples we could take.



the % of sample means away from true mean

tells % of sample

Constructing Confidence Intervals for an Unknown μ and σ known

We never know whether our specific confidence interval covers the true μ or not.

We only know the percent of all CI that could be constructed (with our normality and σ assumptions) that would.

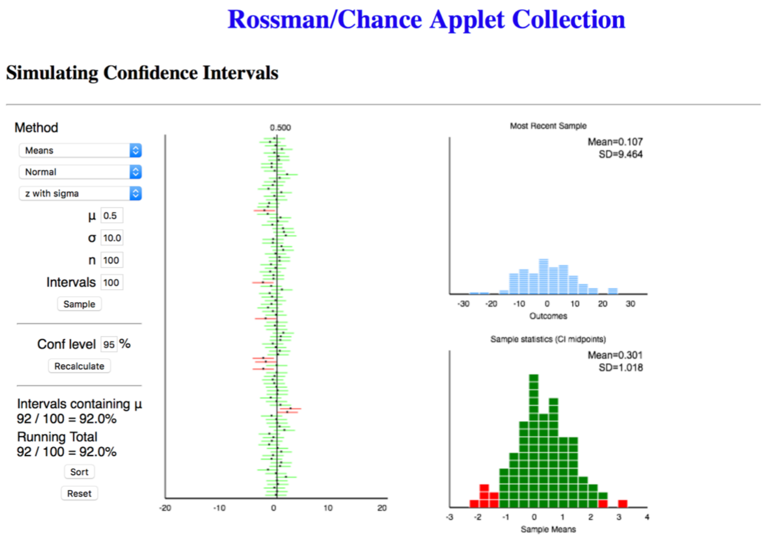
"We are $100\%(1-\alpha)$ confident the interval from (l to u) captures the true parameter

[since it was constructed in such a way that $100\%(1-\alpha)$ of all intervals would]"

Summary:
If $X_1, X_2, \ldots X_n$ is a simple random sample from $N(\mu, \sigma^2)$, then the interval $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ contains μ for a proportion of $1-\alpha$ of random Samples.
Each realization using a sample estimate is a $100\%(1-\alpha)$ confidence interval.

Constructing Confidence Intervals for an Unknown μ and σ known

Simulation: http://www.rossmanchance.com/applets/ConfSim.html



Constructing Confidence Intervals for an Unknown μ and σ known

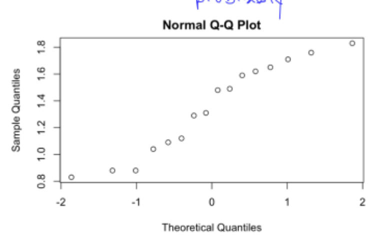
Ex 1 : Suppose we [somehow] know the population sd for all paint thickness: $\sigma_{paint} = 0.3$.
Construct a 95% confidence interval for the population mean paint thickness: μ.

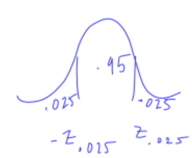
n= 16; Is sample from normal population or is n large enough for the CLT to kick in? so $\bar{X} \sim N$ qqnorm( )
(check qqnorm(paint)) probably

$Z_{\frac{.05}{2}}$ $Z_{.025}$

$1-\alpha=$ 0.95 $\alpha=$ 0.05 $\quad z_{\alpha/2} =$ 1.96

.025 in table





$\bar{x} =$ 1.35 , $\mathrm{SE}(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{.3}{\sqrt{16}} = \frac{.3}{4} = .075$

mean

ME/half width: $1.96(.075) = .147$

95% CI: $\bar{X} \pm Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.35 \pm .147 = (1.203, 1.497)$

We are 95% confident the interval from (1.203 , 1.497) covers the true mean paint thickness.

Confidence Interval Behavior with known σ

CI: $\left(L = \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \mathrm{U} = \bar{X} + z_{\frac{\alpha}{2}} \boxed{\frac{\sigma}{\sqrt{n}}}\right)$

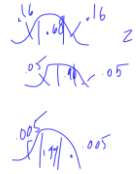
SE($\bar{X}$)

$\bar{X} \pm$ 1 $\frac{\sigma}{\sqrt{n}}$ Is a 68% CI for μ

$\bar{X} \pm$ 1.65 $\frac{\sigma}{\sqrt{n}}$ Is a 90% CI for μ

$\bar{X} \pm$ 1.96 $\frac{\sigma}{\sqrt{n}}$ Is a 95% CI for μ

$\bar{X} \pm$ 2.57 $\frac{\sigma}{\sqrt{n}}$ Is a 99% CI for μ



z .16 below = -.99

Our confidence interval will be wider when:

1. We have an imprecise measure of $\bar{X}$, SE($\bar{X}$) is large
   - n is small
   - σ is large
2. We set a larger confidence level $100\%(1-\alpha)$
   - $z_{\frac{\alpha}{2}}$ is larger

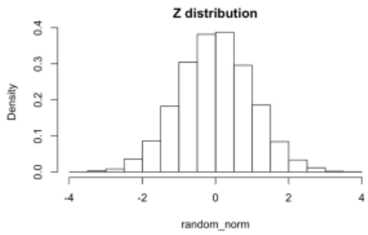
It is always desirable to have the confidence level as high as possible, and the confidence interval as narrow as possible, because these would be indications of a very accurate estimate.

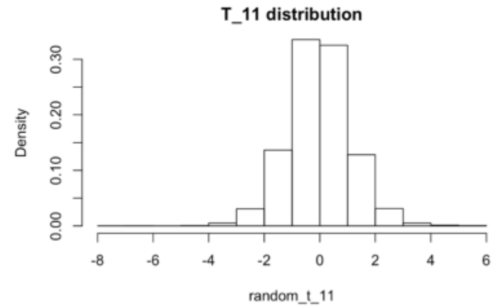
More Realistically, ……Confidence intervals with unknown σ

If we don't have the true population standard deviation σ, we need to estimate it with the sample standard deviation S.

- Where σ is a constant, S is a RV and will be different sample-to-sample
- $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = T_v$ is no longer Normally distributed, it will have a wider/more variable distribution because we are estimating σ with S.
  - How "spread out" the statistic T_v is or how "heavy the tails" are depends on the sample size
  - "Degrees of freedom" determines the spread and is defined $v = n-1$
  - As $v \to \infty$, S better estimates σ and T_v is indistinguishable from Z.
  - When n > 30, T_v is very similar to Z
- Student's t Distribution was was first discovered in 1908 by W. S. Gosset (pseudonym 'Student'), who worked at the time at Guinness Brewing Company, mostly on barley experiments

The T_v Distributions compared to Z

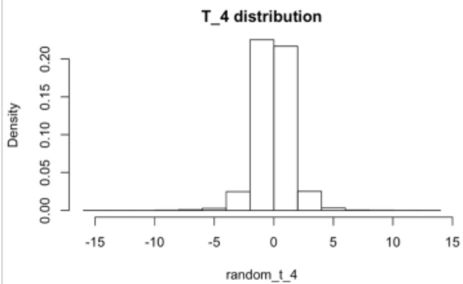


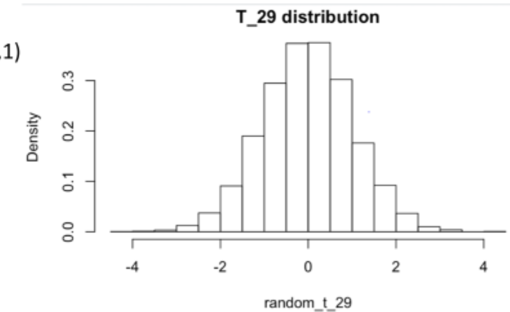


T_{n-1} looks like N(0,1)

*Symmetric about 0
*Single peaked
*Bell-Shaped

As n increases, T_{n-1} is indistinguishable from N(0,1)

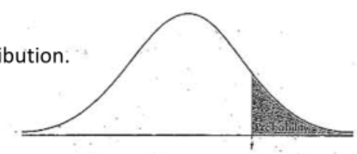




The T_v Distributions

Let $t_{n-1,\propto}$ be the critical value t cutting off an upper area of ∝ from the t_{n-1} distribution.

Student t tables often give upper tail probabilities, using υ "nu" for n-1



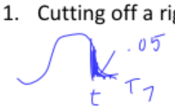
Use the t table to find the critical value t

$n = 8$

TABLE B: t-DISTRIBUTION CRITICAL VALUES

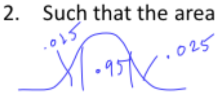
	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	$\boxed{.868}$	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922

1. Cutting off a right tail area of 0.05 from the t_7 distribution $t_{7,0.05}$



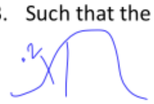
$t_{(7,\ .05)} = 1.895$ $z = 1.65$

2. Such that the area under the t_{18} curve between –t and t is 95%



$t_{(18,\ .025)} = 2.101$ $z = 1.96$

3. Such that the area under the t_{14} curve left of t is 0.20



$t_{(.2,\ t_{(14,.2)})} = .868$ $t_{(14,\ .8)} = -.868$

4. How do these values compare to the same z critical value?

exam ?

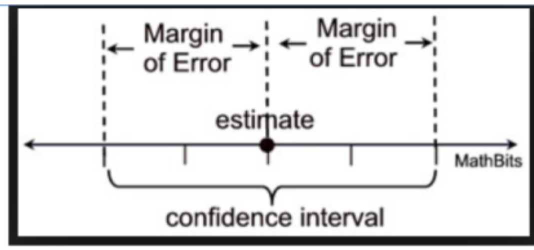
T-table is wider and more distributed because we are estimating sigma with S

Constructing Confidence Intervals for an Unknown μ and σ

Summary:
If $X_1, X_2, \ldots X_n$ is a simple random sample from $N(\mu, \sigma^2)$ (or where n is large enough for the CLT to apply) then the interval $\bar{X} \pm t_{n-1, \frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ contains μ for a proportion of $1-\alpha$ of random samples.

$$\mathrm{P}\left(\bar{X} - t_{\left(n-1, \frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\left(n-1, \frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Each realization using a sample estimate is a $100\%(1-\alpha)$ confidence interval.



This is the $100\%(1 - \alpha)$ t confidence interval for μ, useful when we don't know σ and have a sample of any size from a normal population or a large sample from (almost) any population.

Constructing Confidence Intervals for an Unknown μ and σ

true μ paint thickness

Ex 1: Make a 95% confidence interval for the paint data (assuming we don't know σ).

is population normal? qqplot of sample looked ok $n=16$

$$\bar{x} \pm t_{(15,\ .025)} \frac{s}{\sqrt{n}}$$

mean(paint) = 1.35
sd(paint) = .339

$$1.35 \pm 2.131\left(\frac{.339}{\sqrt{16}}\right) = 1.35 \pm 0.181 = (1.169,\ 1.531)$$

We are 95% confident the interval from (1.17, 1.53) covers the true mean paint thickness.

How does this compare to our z interval? Why does that make sense?

$z \approx (1.203,\ 1.497)$
same center but z's narrower
because $\sigma = .3$ (smaller)
$z_{.025} =$ (smaller)

Choosing a Sample Size with known (or estimated) σ (OL sec 5.3)

Suppose we desire a more precise (narrower) confidence interval, but we want to keep the same confidence level $1-\alpha$?

For a fixed confidence level $1-\alpha$ and an assumed σ, we can use the margin of error to determine the minimum sample size n. *(use $\hat{\sigma} = s$ in usual case that σ is unknown.)*

Ex 1: What sample size is required to reduce the error margin[half width] of the paint thickness 95% confidence interval, above, to 0.1 mil assuming $\hat{\sigma} = s = 0.339$?

$$ME = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \approx z_{\frac{\alpha}{2}} \frac{.339}{\sqrt{n}} = 1.96\left(\frac{.339}{\sqrt{n}}\right) = .1 \qquad \left(\frac{1.96(.339)}{.1}\right)^2 = n$$

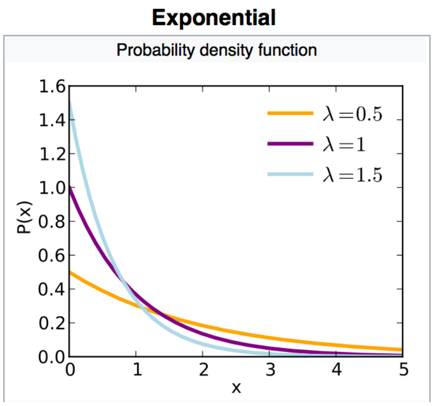
95% → 1.96

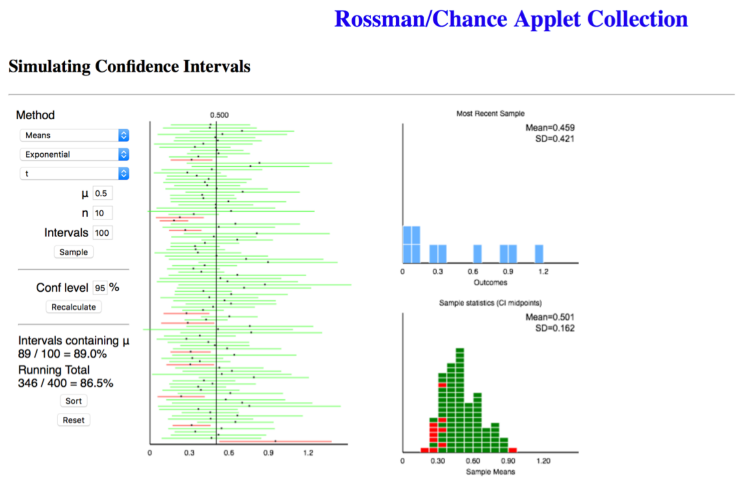
45 observations = 44.15 $n = 44.15$

Will a confidence interval constructed with sample size of 45 observations necessarily capture the true population mean? no

Constructing Confidence Intervals for an Unknown μ, σ and non Normal Pop

Simulation: http://www.rossmanchance.com/applets/ConfSim.html





Constructing Confidence Intervals for an Unknown μ and σ

we dont know if pop is normal, hope n=50 is big enough for CLT to make $\bar{X} \sim N$

Ex 2: A credit company randomly selected 50 contested items and recorded the dollar amount being contested. These contested items had sample mean $\bar{x}$ = \$95.74 and s= \$24.63.

a. Construct a 90% confidence interval for μ. Interpret it in context.

49 Ttable round down

$\bar{X} \pm t_{(49,.05)} \frac{s}{\sqrt{n}} = 95.74 \pm 1.684 \left(\frac{24.63}{\sqrt{50}}\right)$

$= (89.87, 101.61)$

.05 .90 .05

b. Construct a 99% confidence interval for μ. Interpret it in context.

$95.74 \pm 2.704 \left(\frac{24.63}{\sqrt{50}}\right) = (86.32, 105.16)$

.005 .99 .005

$t_{(49,.005)}$

$40 \to .005$

99% confident the interval (86.32, 105.16) captures the true mean \$ amount using contested

Extra Practice:

$\mu = 4.2$

$\sigma = 1.4$

Ex 3: The basal diameter of a sea anemone indicates its age. The population of anemone at the Boston aquarium has a mean diameter of 4.2 cm with standard deviation of 1.4cm.

a. Describe how you would find the probability a randomly-chosen anemone has diameter less than 4 cm.

maybe find area/percentage to left of 4 in histogram

b. Suppose the population of sea anemone diameters are approximately normal. Find the same probability.

4.2

$P(D<4) = P\left(z < \frac{4-4.2}{1.4}\right) = .443$

one tail and assume pop is normal

$D \sim N(4.2, 1.4^2)$

4.2

$P(D<4$

c. Drop the supposition from b and suppose a simple random sample of 40 anemones is taken. Find the probability that the sample mean diameter is less than 4 cm.

dont assume normal

$\bar{D} \approx N\left(4.2, \frac{(1.4)^2}{40}\right)$ $P(\bar{D}<4) = P\left(z < \frac{4-4.2}{1.4/\sqrt{40}}\right)$

CLT

central limit theorem

$P(z < -.935) = 0.183$

$D \sim ?(4.2, 1.4^2)$

$n=40$

$P(\bar{D}<4) = P\left(z < \frac{4-4.2}{\sqrt{\frac{1.4^2}{40}}}\right)$

σ

4 4.2

CLT

$\bar{D} \sim N\left(4.2, \frac{1.4^2}{40}\right)$

Extra Practice:

Ex 3: Now suppose the population mean and standard deviation for the anemones diameter is unknown at the Shedd Aquarium.

d. Here are the diameters of a simple random sample of 40 anemones: 4.3, 5.7, 3.9, 4.8, 3.5, 3.5, 1.3, 4.6, 4.4, 3.7, 4.9, 5.6, 5.1, 2.3, 2.3, 6.9, 5.4, 3.6, 4.3, 4.1, 3.2, 4.6, 2.8, 4.9, 4.5, 4.4, 5.8, 3.6, 5.6, 2.6, 1.5, 4.1, 4.7, 6.5, 5.4, 3.8, 3.4, 4.9, 5.5, 7.2. These data have $\bar{x}$ = 4.33 and s = 1.329. Find a 95% confidence interval for μ or explain why you cannot.

standard error $\frac{\sigma^2}{n}$

qq plot looks linear, so plausible sample from Normal pop. but $n=40$

reassurance we CLT ensures $\bar{X} \sim N$

$\bar{x} + t_{(.025, 39)}\left(\frac{1.329}{\sqrt{40}}\right) = 4.33 + 2.042\left(\frac{1.329}{\sqrt{40}}\right) = (3.9, 4.8)$

(L, U)

standard deviation sample data

3, 7, 9, 12, 15, 18

$\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

only when have whole pop info

same thing

$(x-E(x))^2 * P(X=x) + \ldots$

Identify each of the following interpretations of the above 95% CI Yes, No or Cannot Tell

i. The population mean will lie in the interval from d. F

ii. The sample mean will lie in the interval from d. T

iii. In a future sample of 40 anemones, the sample mean will fall in the interval from d. Cant tell

iv. 95% of the sample anemone ~~weights~~ diameters lie in the interval from d. cant tell

v. 95% of the population anemone ~~weights~~ diameters lie in the interval from d. F

$\bar{x} - z\frac{\sigma}{\sqrt{n}}$ μ $\bar{x} + z\frac{\sigma}{\sqrt{n}}$

$\hat{\mu}$ = estimate

Extra Practice:

t.test()

conf.level = .95

e. Here is a simple random sample of 12 anemone diameters: 5.3, 2.8, 5.2, 2.9, 2.5, 2.9, 3.0, 2.9, 5.2, 4.3, 3.7, 2.7. Find a 95% confidence interval for μ or explain why you cannot.

* check qq plot cant do it

f. Here is a simple random sample of 12 anemone diameters: 3.5, 6.5, 3.6, 2.8, 4.2, 4.2, 1.8, 5.7, 2.6, 4.7, 4.9, 4.4. Find a 95% confidence interval for μ or explain why you cannot.

$\bar{x} = 4.075$ $s = 1.323$ $4.075 \pm t_{(.025, 11)}\frac{1.323}{\sqrt{12}} = 4.075 \pm 2.201\left(\frac{1.323}{\sqrt{12}}\right)$

$= (3.24, 4.91)$

g. Using the standard deviation from part c as an estimate for σ, how large of a sample size would we need so a 95% confidence interval will have a total width less than 0.5 units.

$.25 = z_{.025}\frac{1.323}{\sqrt{n}}$ $.25 = 1.96\left(\frac{1.323}{\sqrt{n}}\right)$ $n = 108$

Ex 4: In general, how does doubling the sample size change the confidence interval width? (For a z interval)

$\frac{(\text{mult})\sigma}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\frac{\text{mult}\,\sigma}{n}$ smaller by factor $\frac{1}{\sqrt{2}}$

For Next Time

- Start/Continue working through posted homework 3. Post questions on Piazza.

- I will be posting additional Exam 1 practice questions (~ Thurs) which we will have some class time to work through/answer questions on next Tuesday.

- Continue working on Quiz 2 due Tues Oct 9th