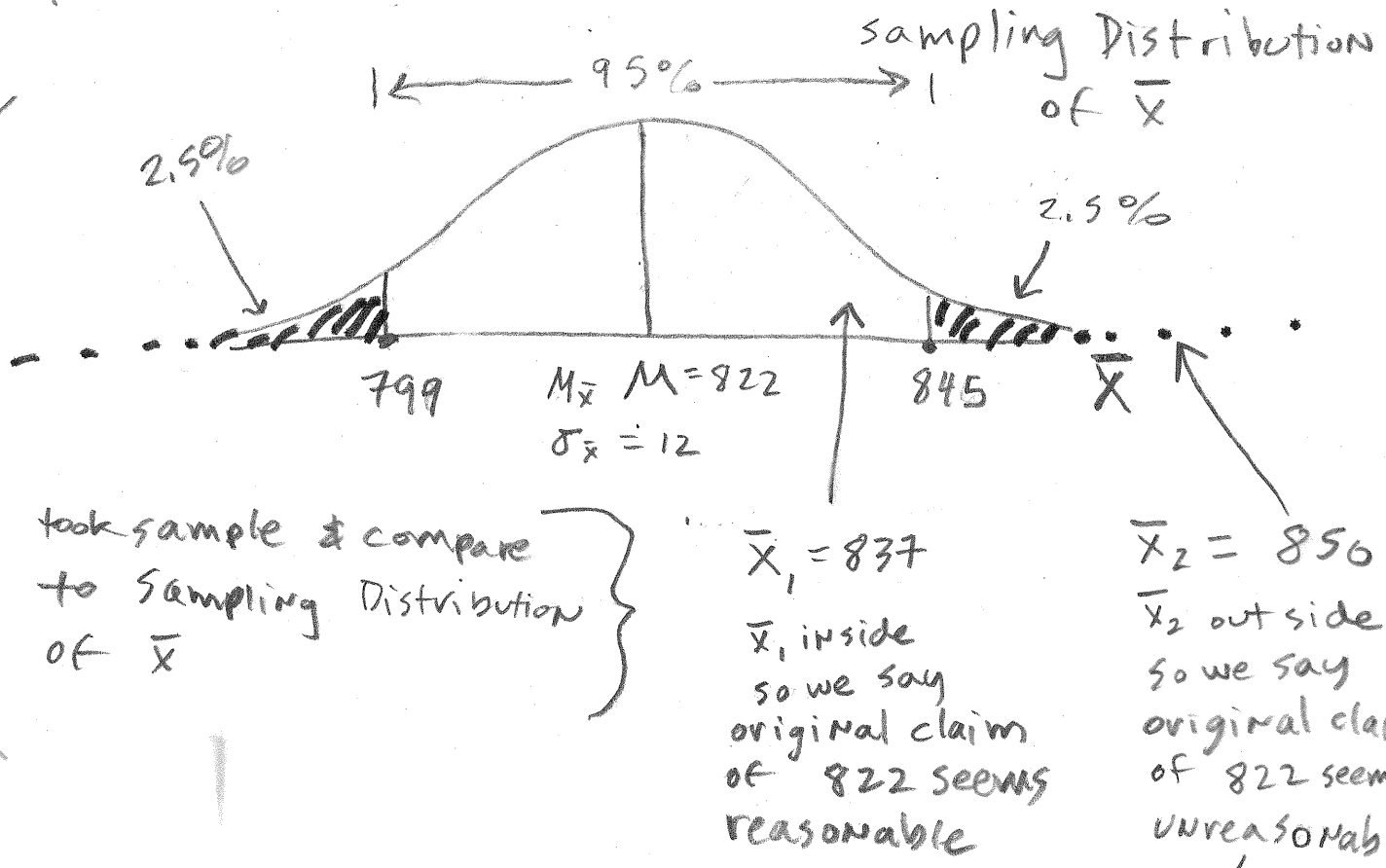
Chapter 7: Know Pop. Mean, M — SDO $\bar{x}$

Ch. 7



(1) Even though 2.5% of samples would occur out here, because getting a sample of 850 is so unlikely, we say that from our sample evidence it seems unlikely, and thus original claim seems unreasonable.

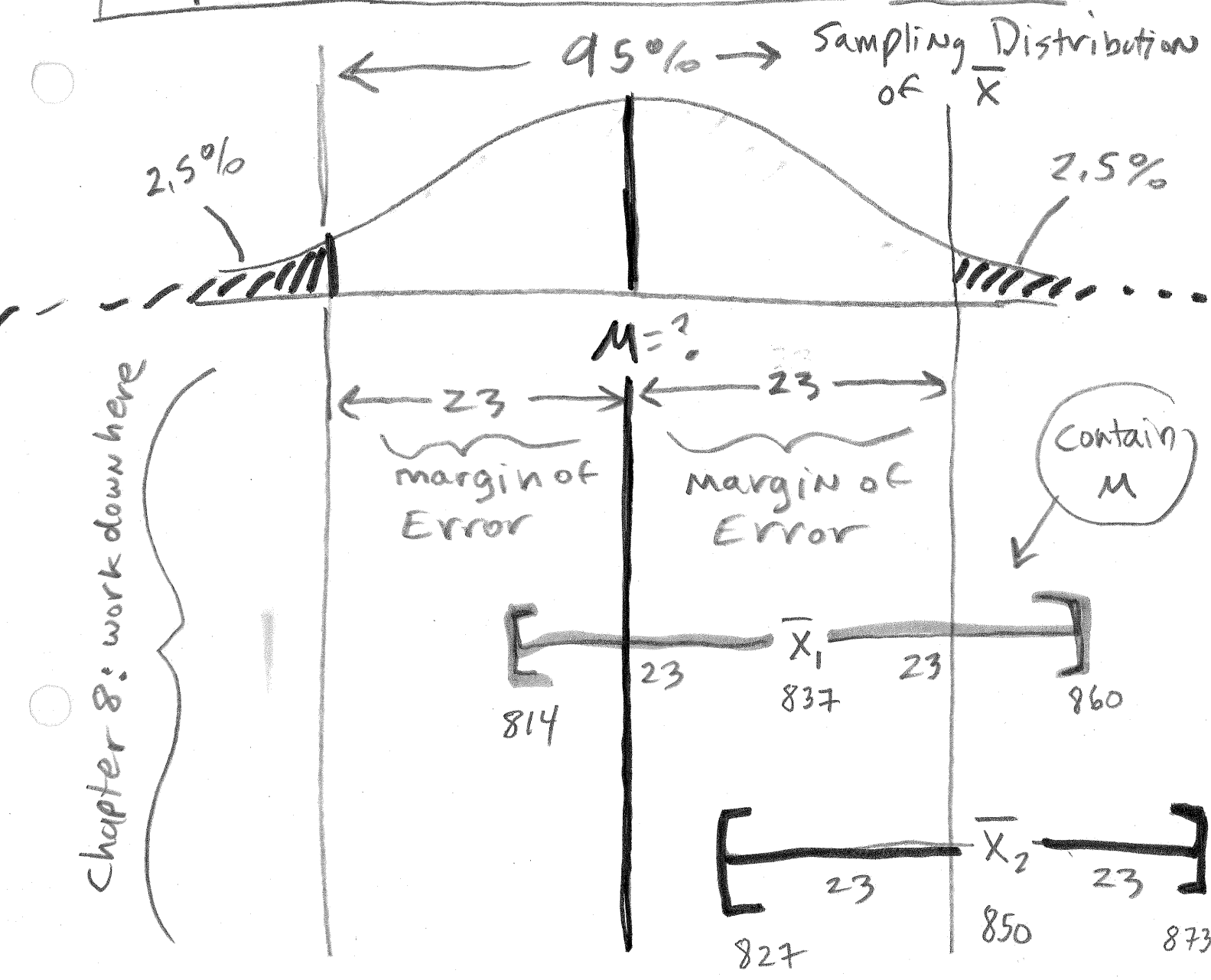
(2) Because samples will almost always have sampling Error, the sampling Error with $\bar{x}$ = 850 is Not acceptable, whereas with $\bar{x}$ = 837 the sampling Error is acceptable.

(3) we create point estimates & compare them directly to SDO $\bar{x}$.

(4) we make statements like: we are 95% sure that $\bar{x}$ will lie between 799 & 845

chapter 8: Don't Know Pop Mean, M

SDo $\bar{x}$ (ch.8)



① we run the risk that 2.5% of our intervals will not contain M & will be on upper side.

② we develope Confidence Intervals to estimate the pop. parameter (we will estimate $\bar{x}$ & $\bar{p}$)

③ we make statements like: we are 95% sure that M will be in our interval

Chapter 8
Confidence Intervals (Interval Estimation)

chapter 7

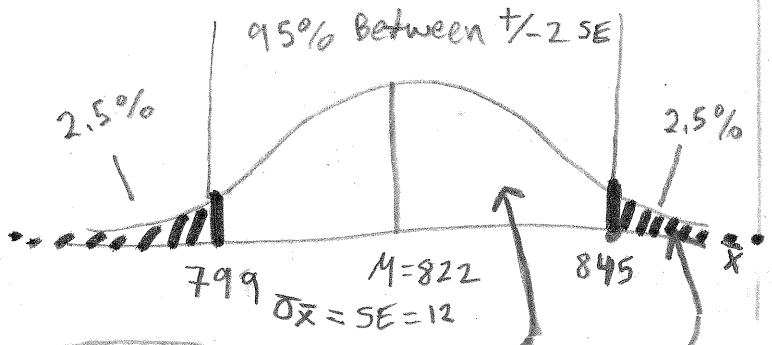
Know what pop. mean, μ, is or what it should be

when insurance company says "average cost for policy is $822, we know what μ should be

↓

we develop point estimates

& compare them directly to Sampling Distribution of $\bar{x}$



If point estimate is $\bar{x}_1$ = $831 original claim seems reasonable

If point est. is $\bar{x}_2$ = $850 original claim seems unreasonable

Make statements about $\bar{x}$:

we are 95% sure that $\bar{x}$ will lie between 799 and 845

chapter 8

Don't know what pop. mean, μ, is or what it should be

when printer cartridge manufacturer wants an "average number of pages per cartridge", they can't calculate pop. mean, μ, because they can't print all pages from all cartridges.

we develop Confidence Intervals

if our point estimate is

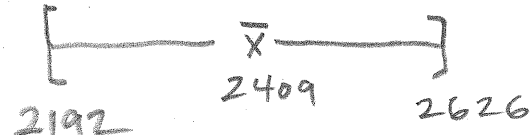
$\bar{x}$ = 2409 pages

we can't compare $\bar{x}$ directly to Sampling Distribution of $\bar{x}$ but we can:

Add a "Margin of Error" to our point estimate:

+/- 217 pages

to create a 95% confidence interval



And say:

we are 95% sure that μ will lie between 2192 & 2626 pages

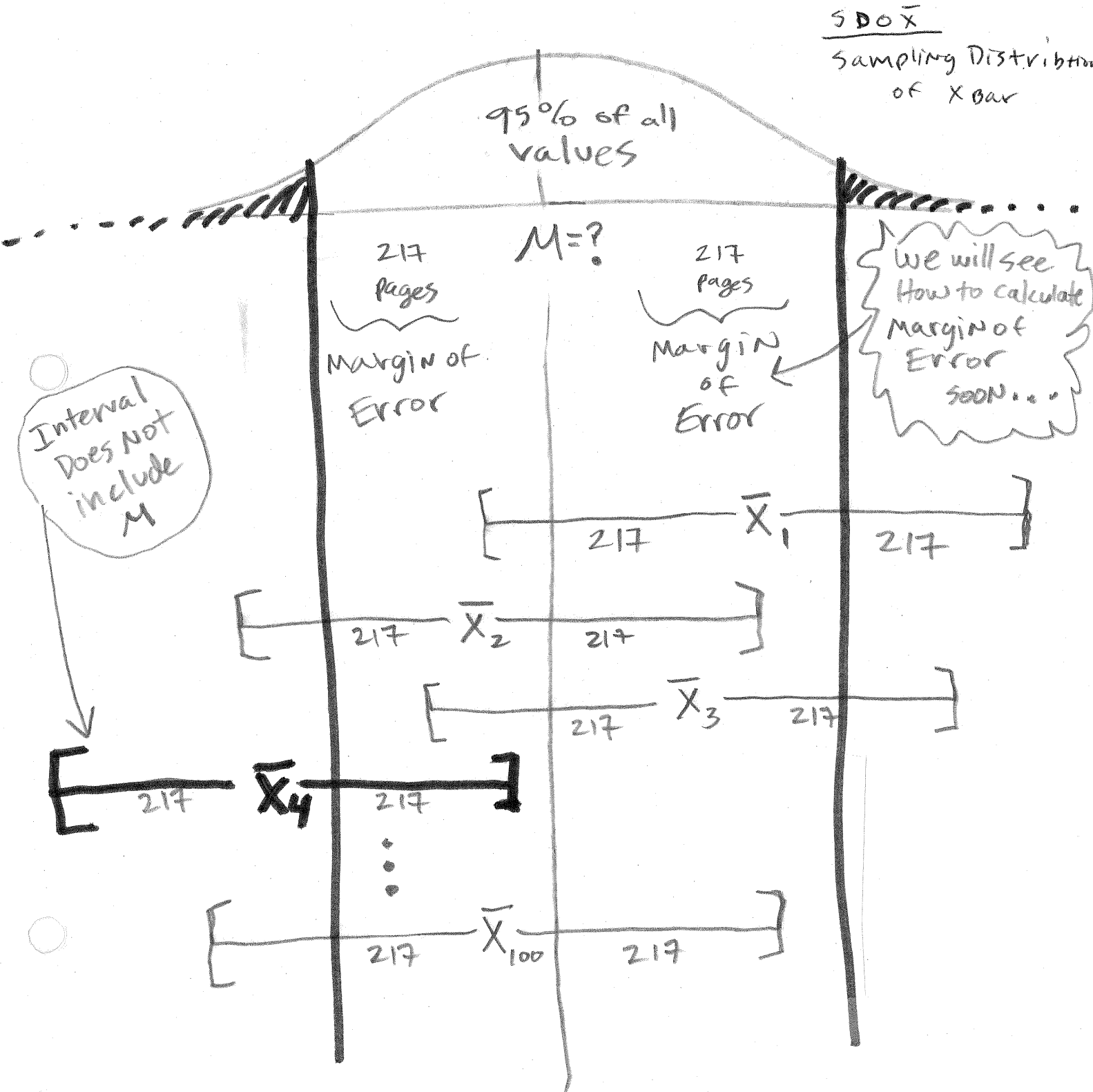
or

Ave pages will be 2409 w/ Margin of Error 217

For 95% Confidence Intervals

Because we don't know M

technically it would look like this:



→ Notice for printer cartridge example:

We CAN NOT calculate population mean, μ.

(why?)

Because we would have to test every cartridge made or count every page from every cartridge in every house...

Other Examples:

(1) we do not KNOW how many meals married couples eat out each week
- we can take a sample and our $\bar{x}$ (mean) will be the point estimate for the unknown pop. mean.

(2) we do not KNOW what the mean amount of profit is per auto sold in USA.
- we can take a sample and our $\bar{x}$ will be the point estimate for the unknown pop. mean.

Since $\bar{x}$ rarely (if ever) equals μ we can't just take sample & use $\bar{x}$, we must add Margin of Error to each side of $\bar{x}$ & create Confidence Interval for our estimate of pop. mean, μ.

Confidence Intervals (Interval Estimation)

① An estimate of the population parameter that provides an interval believed to contain the value of the parameter.

② estimate of population parameter = point estimate +/- Margin of Error

Confidence Level or Confidence Coefficient

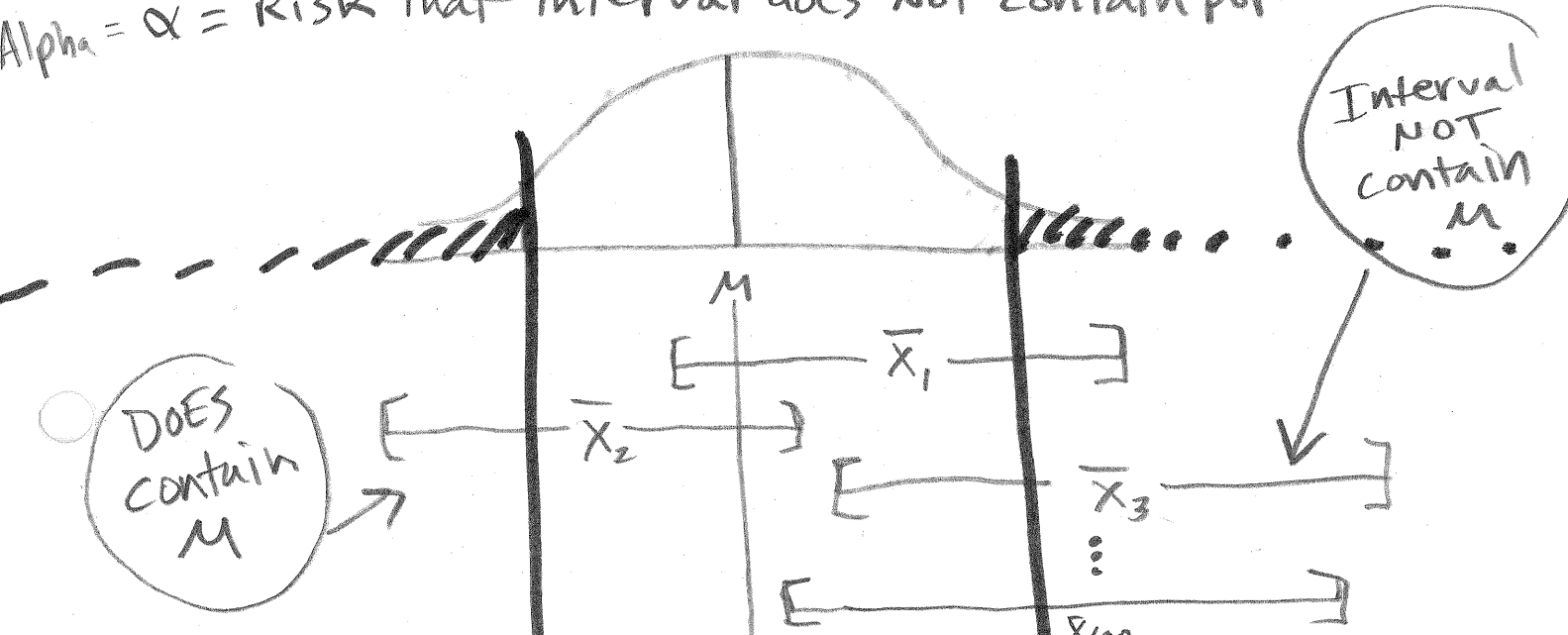
probability that population parameter will occur within our interval.

Examples: 90%
95%
99%

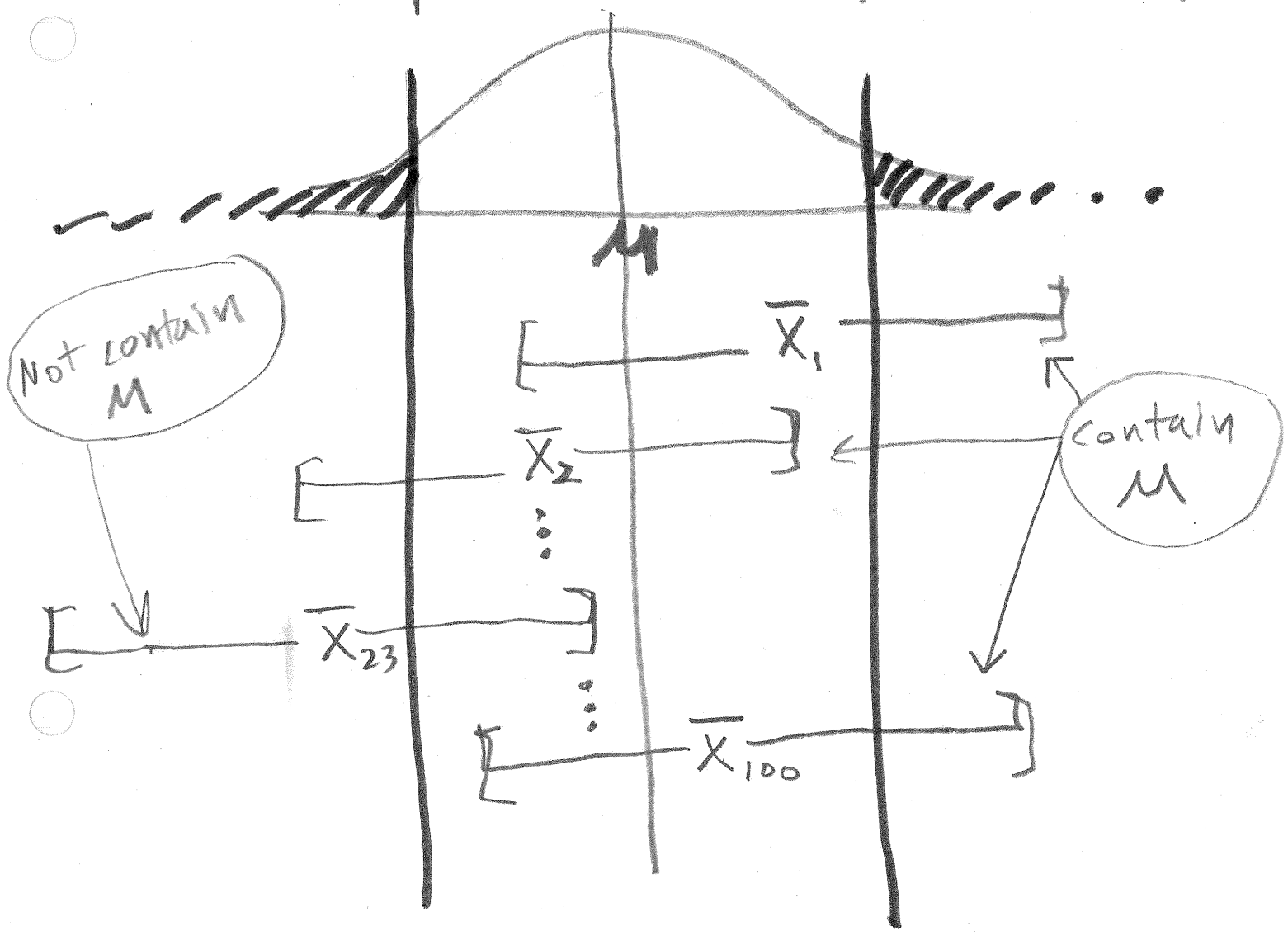
Level of Significance = α = Alpha

Alpha = α = 1 - "Confidence Interval"

Alpha = α = Risk that interval does not contain pop. parameter.



Note: 95% Confidence Interval



★ if population Distribution follows a Normal Distribution, and you construct a 95% confidence Interval, 95 would contain μ, & 5 would NOT contain μ.

★ If pop. Distribution Not Normal, the sample size must be sufficiently big (central Limit Theorem) so that approximately 95 have μ & 5 not have μ.

Confidence Interval for μ σ KNOWN*

*KNOWN or have good estimate

$$\bar{X} \pm \underbrace{Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}}_{\text{Margin of Error}} \quad \left.\right\} \text{Standard Error}$$

σ = sigma = Population Standard Deviation

$\bar{X}$ = sample mean

z = standardize value = number of standard Error

$\sigma_{\bar{x}}$ = Standard Error = Standard Deviation of sampling Distribution of $\bar{x}$

$1-\alpha$ = confidence level (confidence coefficient)

α = alpha = risk μ Not in Interval

Margin of Error = Amount on either side of $\bar{x}$ to create Interval

$Z_{\alpha/2}$ = z on upper end

n = sample size

Excel

Calculate Z

Method 1:

upperLimit = NORM.INV$(1-\frac{\alpha}{2}, \bar{X}, \sigma_{\bar{x}})$

Lower Limit = NORM.INV$(\frac{\alpha}{2}, \bar{X}\ \sigma_{\bar{x}})$

method 2: Z_{upper} = NORM.S.INV$(1-\frac{\alpha}{2})$

Calculate Margin of Error

Method 3:

{Margin of Error} = CONFIDENCE.NORM(α, σ, n)

σ KNOWN or at least reliably estimated

(1) When there is large amounts of historic data, sigma (σ) may be KNOWN.

(2) Quality control applications where process is assumed to be operating correctly, it is appropriate to treat sigma as KNOWN

(3) Most situations σ is NOT KNOWN, so we don't use Normal (z) Distribution, we use t-distribution. In this case we will have to use the same sample to estimate μ & σ. We will use $\bar{x}$ bar & s. (more later....)

Note: $\bar{x} +/- Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$

If particular sample size provides an interval that is too wide for any practical use, increase sample size, n.

$SE = \frac{\sigma}{\sqrt{n}}$ ← As this increases (sample size)

$\frac{\sigma}{\sqrt{n}}$ decreases (Standard Error)

$Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$ decreases (Margin of Error)

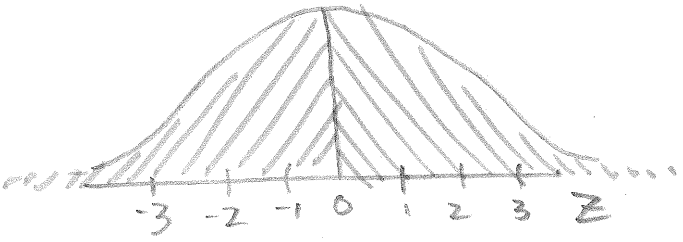
Narrower interval, & thus greater precision of estimate of μ.

Q: But what if we do not KNOW the population standard deviation, sigma (σ)? Can we still construct Confidence Intervals?

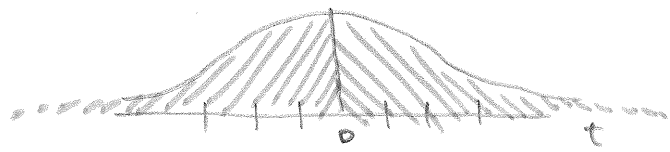
A: Yes we can! But will have to use the point estimate sample standard deviation, s, in place of σ.

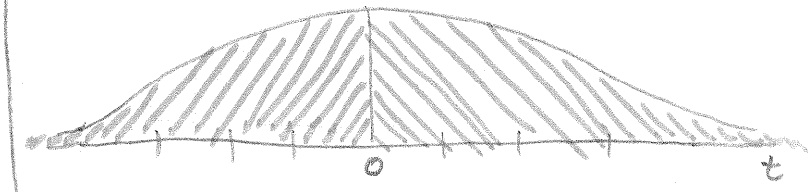
When we use the sample standard deviation, s, instead of the population standard deviation, σ, we cannot use the standard Normal curve & the z-score. When we use the sample standard deviation, s, we must use the t-distribution.

σ KNOWN	σ NOT KNOWN



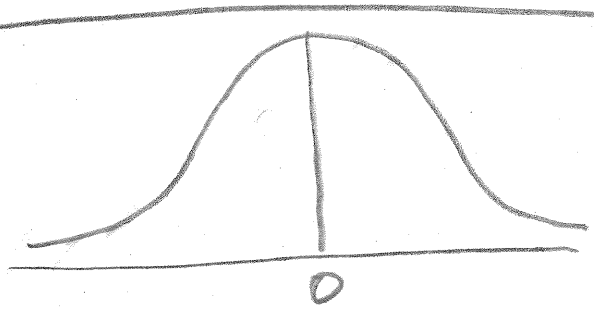
1 standard Normal Curve using z



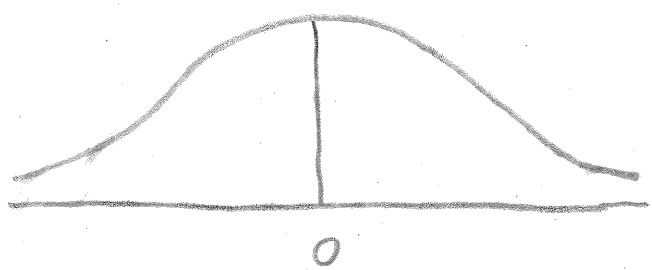


many t-distribution curves

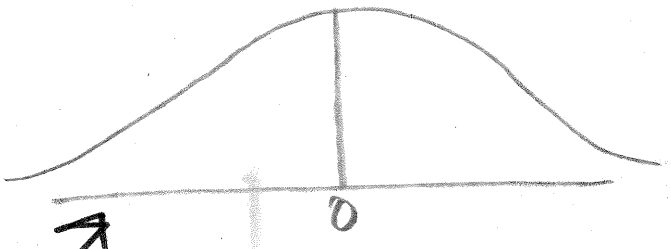
Many t - Distributions



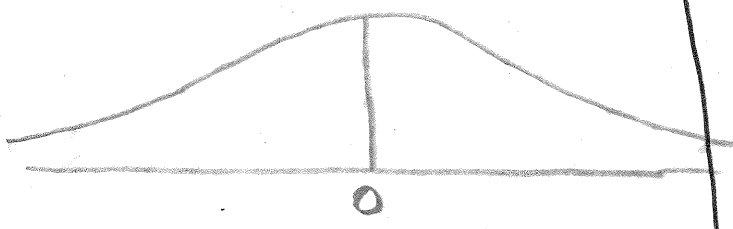
n = 100
df = 99



n = 50
df = 49



n = 30
df = 29



n = 15
df = 14

One for each sample size, n, or Degrees of Freedom, n-1.

① This means that when creating our Confidense Interval, we will have to choose the correct t-Distribution

- If we have n = 50, we choose
- If we have n = 30, we choose

10

Degrees of freedom (statistics)

From Wikipedia, the free encyclopedia

In statistics, the number of **degrees of freedom** is the number of values in the final calculation of a statistic that are free to vary.[1]

Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom (df). In general, the degrees of freedom of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself (which, in sample variance, is one, since the sample mean is the only intermediate step).[2]

Mathematically, degrees of freedom is the dimension of the domain of a random vector, or essentially the number of 'free' components: how many components need to be known before the vector is fully determined.

The term is most often used in the context of linear models (linear regression, analysis of variance), where certain random vectors are constrained to lie in linear subspaces, and the number of degrees of freedom is the dimension of the subspace. The degrees-of-freedom are also commonly associated with the squared lengths (or "Sum of Squares") of such vectors, and the parameters of chi-squared and other distributions that arise in associated statistical testing problems.

While introductory texts may introduce degrees of freedom as distribution parameters or through hypothesis testing, it is the underlying geometry that defines degrees of freedom, and is critical to a proper understanding of the concept. Walker (1940)[3] has stated this succinctly:

> For the person who is unfamiliar with *N*-dimensional geometry or who knows the contributions to modern sampling theory only from secondhand sources such as textbooks, this concept often seems almost mystical, with no practical meaning.

t-Distribution (Student's t-Distribution)

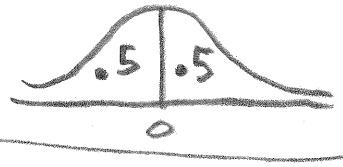
When are we allowed to use t-Distribution?

A: when population Distribution is Normal shaped or near Normal, or n is sufficiently large (Central Limit Theorem)

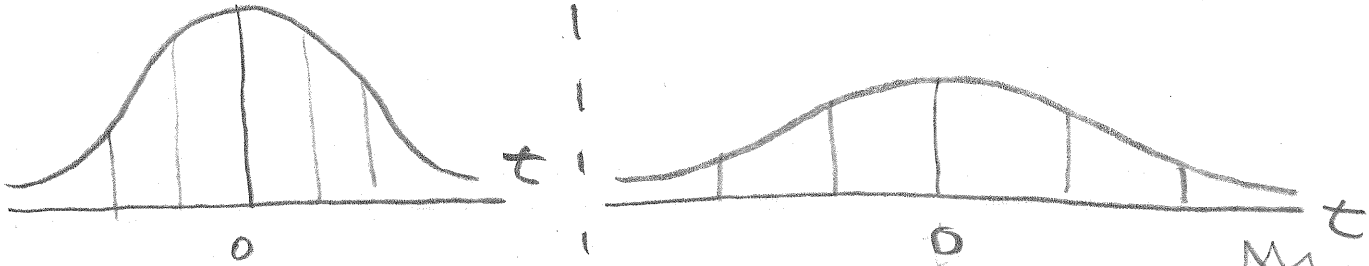
Characteristics of t-Distribution

All Area = 1

1. Continuous Distribution
2. Normal (Bell) shape & symetrical



3. The smaller n or df, the larger the standard Deviation, the flatter the curve,



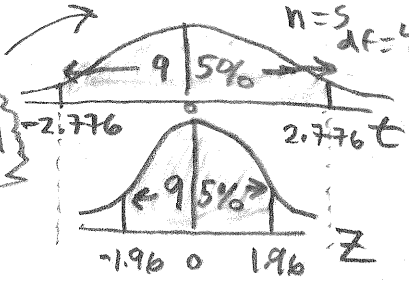
n = 30, df = 29
taller curve
smaller standard Deviation

n = 15, df = 14
Flatter curve
Bigger standard Deviation

Makes Sense: smaller n must make Interval Bigger

4. At center, t-Distribution is Flatter & more spread out than Normal Z Distribution

t will yield Bigger Interval



5. As n or df increase, t-Distribution approaches Z-Distribution.

Confidence Interval for μ σ Not Known

$$\bar{X} +/- \underbrace{t_{\alpha/2} * \frac{s}{\sqrt{n}}}_{\text{Margin of Error}} \quad \left.\right\} \text{Standard Error}$$

$\bar{x}$ = sample mean
s = sample standard Deviation
$1-\alpha$ = confidence level (confidence coefficient)
α = alpha = risk μ not in interval
$t_{\alpha/2}$ = t value on upper end
t = standardize t value = # of standard Errors
n = sample size
Margin of Error = Amount on either side of $\bar{X}$ to create Interval
$s/\sqrt{n}$ = Standard Error = SD of Sampling Distribution of $\bar{x}$
df = Degrees of Freedom = n − # of Samples

Excel

calculate t
Method 1:

t_{upper} = T.INV$(1-\frac{\alpha}{2}, df)$
t_{lower} = T.INV$(\frac{\alpha}{2}, df)$

calculate M.o.E
Method 2:

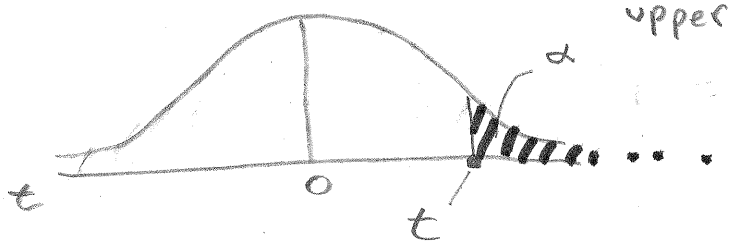
{Margin of Error} = CONFIDENCE.T(α, s, n)

calculate $\bar{X}$, s, M.o.E
Method 3:

{$\bar{X}$ & s and Margin of Error} Data Analysis, Descriptive statistics (Data Ribbon Tab)

Excel t functions σ NOT KNOWN

1 tail to right



upper t = T.INV(1−α, df)

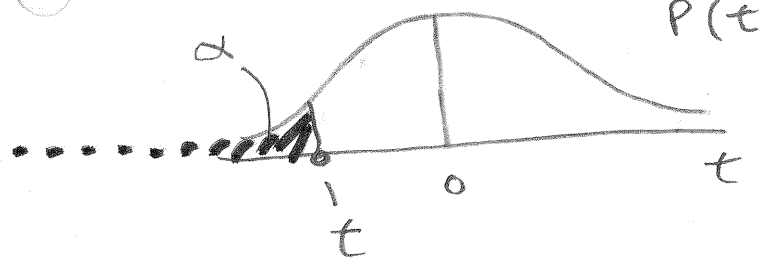
P(t or greater) = T.DIST.RT(t, df)

Neg. or pos. t will give Prob. to Right

OR

1 − T.DIST(t, df, 1)

1 tail to left



lower t = T.INV(α, df)

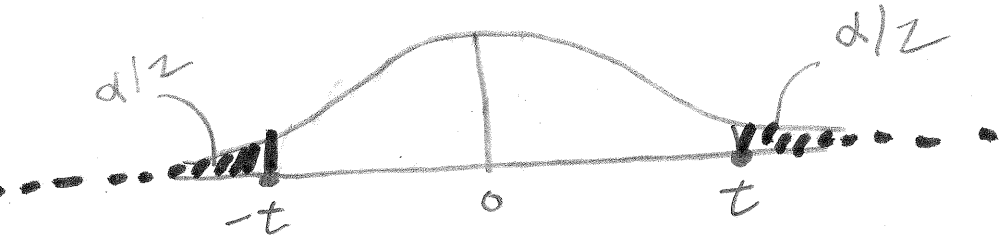
P(t or less) = T.DIST(t, df, 1)

Neg. or pos. will give Prob. to Left

1 = probability

0 = height of curve (used for charting)

2 tail



upper t = T.INV.2T(α, df)

P(t or greater / −t or less) = T.DIST.2T(upper t, df)

P(Between −t & t) = T.DIST(t, df, 1) − T.DIST(−t, df, 1)

Charting t Distributions

$$\left\{\begin{matrix}\text{Height of} \\ \text{curve}\end{matrix}\right\} = T.DIST(t, df, 0)$$

1 = cumulative prob. from $-\infty$ to t

0 = Height of curve

Histograms from Sample Data

1. For a clue about shape of Population Distribution
2. Not conclusive, but sometimes it is best info we have.
3. Helps to determine sample size & whether you can use t distribution.

Note: Create of t-distribution

- created by William Sealy Gosset
- Gosset wrote under name "Student"
- Gosset was OXford graduate in Mathematics
- Gosset worked for Guinness Brewery (Dublin, Ireland)
- Developed t Distribution while working on small scale materials & temperature experiments
- Gosset's mathematical development of t distribution assumed that populations had Normal Distributions.
- Research shows that t distribution can be applied in some situation where pop. Distribution deviates from Normal distribution

Guidelines for using t distributions when pop Not Normal.

① If pop. IS Normal, if you create 100 similar 95% Confidence Intervals, 95 should contain pop. μ & 5 not.

② If pop Not Normal, if you create 100 similar 95% Confidence Intervals, about 95 should contain μ & about 5 will not if:

- Art form
- Professional Judgement.

- Distribution almost symetrical $n \geq 15$
- Distribution Near Normal $n \geq 30$
- Highly skewed or has outliers $n \geq 50$

Printers Manufacturer

Printers Manufacturer is a manufacturer of ink jet printers. They would like to include as part of their advertising the number of pages a user can expect from an ink cartridge. A sample of 10 cartridges was taken with a mean of 2409 pages and a standard deviation of 304 pages.

8. List variables:

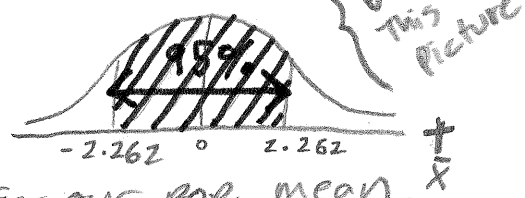
$n = 10$

$df = 10 - 1 = 9$

$\bar{X} = 2409$ pages

$s = 304$ pages

$\frac{s}{\sqrt{n}} = \frac{304}{\sqrt{10}} = 96.13$



our interval BELOW this picture

9. What is the best estimation for the population mean?

$\bar{X} = 2409$ pages is the best estimate for our pop. mean. "point estimate"

10. Determine a 95% Confidence Interval

a. State the level of confidence, state the df, then look up t in back of book.

Level of confidence = .95 $df = 10 - 1 = 9$ $\Rightarrow t = 2.262$

Excel = TINV(1-.95, 9)

2409 ± 217.45 (Margin of error)

b. Using the correct confidence interval formula, calculate the confidence limits

$\bar{x} \pm t \frac{s}{\sqrt{n}}$

$2409 \pm 2.262 * 96.13$

2192 & 2626

"The limits for the 95% CI are 2192 pages and 2626 pages.

11. Conclusions:

- The customer can expect an average of 2409 pages/cartridge.
- The typical usage ranges from 2192 pages to 2626 pages (given a 95% CI). The margin of error is .
- The margin of error is 217.45 pages.

We are 95% sure that the population mean lies between 2192 & 2626 pages.

- If we constructed 100 similar intervals, 95 of them would contain pop. mean.

Proportions

Proportions: The fraction, ratio, or percent indicating the part of the sample or the population having a particular trait of interest.

Example: A recent survey of Highline students indicated that 98 out of 100 surveyed thought textbooks were too expensive.

x = # of successes = particular trait of interest = 98

n = sample size = 100

$\bar{p} = \frac{x}{n}$ = sample proportion

Notes about proportions:

1. The sample proportion is our best estimate of our population proportion, π.
2. Proportions are Nominal Level data.
3. success or failure sounds Binomial, right? so...

↓

In order to build a Confidence Interval for proportion:

must verify:

1. Are there a fixed # of trials?
2. Are results independent?
3. Does each trial result in success or failure?
4. π stay same each trial?
5. $n * \pi > 5$ and $n * (1-\pi) > 5$

construct of confidence Interval for Proportion

X = # of successes

n = sample size

$\frac{x}{n} = \bar{p}$ = sample proportion = best estimate for π or p

alternative Notation

$$\left\{\begin{matrix}\text{Confidence} \\ \text{Limits}\end{matrix}\right\} = \bar{p} \pm Z\sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$$

Margin of error

$p = \pi$ = population proportion

sampling error = $\bar{p} - \pi$ or $\bar{p} - p$

$\bar{p}$ = point estimate

or Books Notations:

$$\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}*(1-\bar{p})}{n}}$$

$\alpha/2$ ↑ area on upper side.

Excel confidence Interval for Proportions

{sample size n} = COUNTA(Range)

{count successes} = COUNTIF(Range, criteria)

{Z value} = NORM.S.INV(1 − α/2)

{standard Error} = SQRT($\bar{P}$*(1-$\bar{P}$)/n)

Furniture Land South

Furniture Land South surveyed their customers (n = 600) to see if they liked the new line of durable foam furniture decorated in bright colors. 414 said they were excited about the new line. All the binomial tests are met.

1. List variables:

$n = 600$
$x = 414$
best estimate for π: $p = \frac{x}{n} = \frac{414}{600} = .69$ → use p for π
Success = excited
Failure = not excited

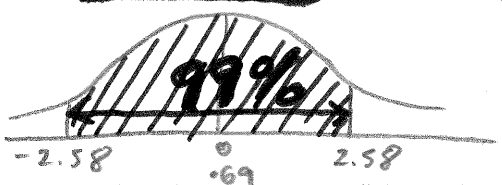
1 Fixed # trials = yes $n = 600$
2 independent ✓
3 S/F ✓
4 $\pi = .69$ same each time ✓
5 $\pi * n > 5 \Rightarrow .69 * 600 = 414 > 5$ ✓
6 $n * (1 - \pi) > 5 \Rightarrow 600 * .31 = 186$ ✓
we can use z!

2. What is the best estimation for the population proportion?

The point estimate $p = .69$ is the best estimate for our population proportion

3. Determine a 99% Confidence Interval
   a. State the level of confidence, [redacted] then find z

Level of confidence = .99 $.99/2 = .495$
$z = \pm 2.58$
{our interval BELOW this picture



   b. Using the correct confidence interval formula, calculate the confidence limits

$p \pm z\sqrt{\frac{p*(1-p)}{n}} = .69 \pm 2.58\sqrt{\frac{.69*.31}{600}} \Rightarrow .69 \pm .0487$

.6413 and .7387 are the confidence limits.

4. Conclusions:

- The owner can be 99% sure that the population proportion (% of customers excited about new product) is between .6413 and .7387. The new product will probably be popular.

- .0487 is Margin of error

- 99 of 100 similarly constructed intervals would contain pop. proportion.

Confidence Interval:

Point estimator +/- Margin of Error

↓

If we know this we can solve for n, sample size.

Point Estimator +/- Error (E)

$$E = Z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

$$\sqrt{n} = \frac{Z_{\alpha/2} * \sigma}{E}$$

Sample size for Interval Estimation for pop. mean

$$n = \left(\frac{Z_{\alpha/2} * \sigma}{E}\right)^2$$

These are both estimates. Judgment must be used

Sample size for Interval Estimation for pop. proportion

$$n = p*(1-p)*\left(\frac{Z_{\alpha/2}}{E}\right)^2$$

{sample size for confidence Interval pop mean} $N = \left(\frac{Z_{\alpha/2} * \sigma}{E} \right)^2$

Requires estimate: ways to get estimate

---

(1) use estimate for sigma from previous studies (or gov. studies)

(2) pilot study before you run experiment

(3) estimate large & small values:

$$\frac{\text{large} - \text{small}}{4} = \text{aproximate } \sigma$$

* Always Round up when calculating n !!

Excel: = ROUNDUP (n estimate, 0)

sample size for Confidence Interval pop proportion

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 * p * (1-p)$$

Requires estimate

How to estimate:

1. use estimate from previous studies or available data or gov. data
2. Pilot study before you run experiment
3. Best guess (requires judgment)
4. None of these apply use $p = .50$

* ($.5*(1-.5) = .25$ and as p increases it will never get bigger than .25). This will always yield largest sample size.

Note: Margin of Error for proportion

* margin of error for estimating pop. proportion is almost always less than 0.10. Gallup & Harris commonly use $E = .03$ or $E = .04$