DA503 Applied Statistics

Lecture 05

Interval Estimation

Problems with point estimators

- Point estimator is a single number that estimates the population parameter varies from sample to sample. So it's almost guaranteed that we'll miss the actual true mean for the population by some, but hopefully not much.
- So we don't have a measure of how certain or confident we are that we actually got the true mean. We need a confidence interval in which the estimator has a likelihood of hitting the true mean, such as:

> We're 95% confident that the true population parameter (say $\boldsymbol{\mu}$) lies within our estimate interval $(\bar{\mathbf{x}}-\boldsymbol{\epsilon} , \bar{\mathbf{x}}+\boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon}$ is some margin of error.

which leads us to the interval estimators.

Introduction

- Till now, we've considered several point estimators.
- We concluded that $\bar{X}$ was a good estimator of μ for populations that were approximately normal.
- Even though on average $\bar{X}$ is on target μ, specific $\bar{X}$ observed is almost certain to be a bit high or a bit low. To be reasonably confident that our inference is right, we cannot claim that μ is exactly equal to the observed $\bar{X}$. So instead:
  - We construct an internal estimate (a Confidence Interval) for the population parameter in the form of:

sampling error = $\bar{x}$ - μ (uncertainty due to variation from one sample to another)

  - **Critical question**: How wide should this allowance or sampling error be?

What is a Confidence Interval (CI)?

POLITICS

Health Care Law

Voters See Worse Care, Higher Costs Under Single-Payer Health System

Wednesday, August 07, 2019

A number of the top contenders for the 2020 Democratic presidential nomination are championing a government-run, single-payer health care system, but voter support is down. Perhaps that's because voters see the quality of care suffering, while their personal costs go up.

The survey of 1,000 Likely Voters was conducted on August 4-5, 2019 by Rasmussen Reports. The margin of sampling error is +/- 3 percentage points with a 95% level of confidence. Field work for all Rasmussen Reports surveys is conducted by Pulse Opinion Research, LLC. See methodology.

This tells us about how certain (or uncertain) we are about the true figure in the population (the width of the confidence interval).

If the sampling were repeated over and over again, the results would match the results from the actual population 95 percent of the time.

What is a Confidence Interval (CI)?

- CI is used to describe the amount of uncertainty associated with a point (sample) estimate of a population parameter (accounting for the variance of the distribution of mean).
- **Interpretation of CI** : Suppose that a 90% CI states that the population mean is greater than 10 and less than 20. How would you interpret this statement?
  - There is a 90% probability that the population mean falls between 10 and 20. Correct or Incorrect?
  - This is not correct! Population parameters are constants, not random variables. So the probability that a constant falls within a certain range is always 0 or 1.
- CI's don't allow one to make probability statements. CI indicates a property of the procedure as is typical for a frequentist technique.

What is a Confidence Interval (cont'd) ?

- **Example**: Your new experimental drug reduces the average length of a cold by 36 hours with a 95% confidence interval between 24 and 48 hours.
  - If you run 100 identical experiments, about 95 of the confidence intervals will contain the true value you're trying to measure.
- Misinterpretation of a confidence interval:

✗ *"There is a 95% probability that μ [the population mean] lies in the confidence interval [24, 48]"*

✓ *"The process by which the interval [24, 48] is computed yields intervals which include μ 95% of the time"*

✓ *"A 95% CI indicates that 95 out of 100 samples from the same population willl produce CI's that contain the true value of the population parameter"*

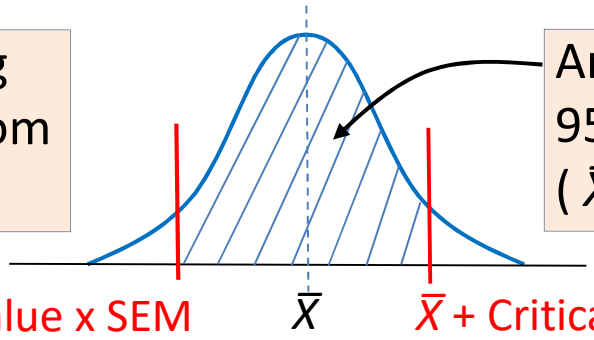
What is a Confidence Interval (cont'd) ?

- The procedure for constructing a, say 95% CI has the property that, over a vast number of study repetitions, 95% of the intervals it generates would contain the true population parameter if the model used was correct.
- The notion of study repetitions is the key to this definition. It means that the definition doesn't apply to any one interval alone, but instead must refer to the intervals produced by the procedure over all possible study repetitions.
- If you are refering to a single CI constructed from a sample already formed, the chance that the one computed interval actually contains the true value is either 1 or 0, i.e., it either contains the true value or not!

How to construct a Confidence Interval

- We need three pieces of information:
  1. **Statistic**: Population parameter to be estimated (sample mean, sample proportion)
  2. **Significance level: α**
     Confidence level (1-α): Percentage of all possible samples that can be expected to include the true population parameter (usually 90%, 95%, 99%).
  3. **CI = Sample statistic ± Margin of error**

A statistic expressing the amount of random **sampling error**



Margin of error = Critical value x SEM

Constructing a CI for the sample mean

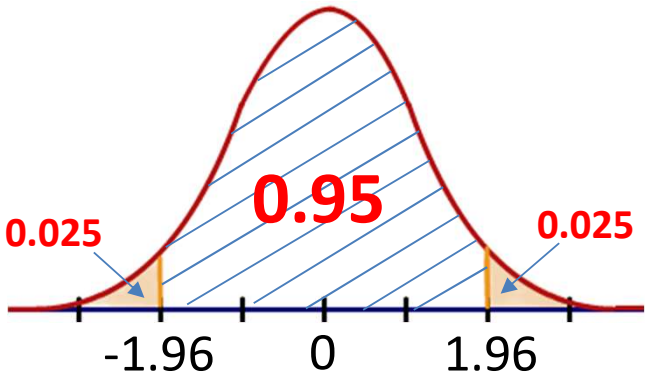
- Assume a sampling from a normally distributed population with a known standard deviation σ.
- The CI for $\bar{X}$ will be: $\bar{X}$ ± Margin of error , where $\bar{X}$ is a normally distributed random variable with a mean μ of the sampling distribution and a standard deviation of the sample means (SE of the sampling distribution) $\sigma_{\bar{x}} = \sigma / \sqrt{N}$

If $\bar{X}$ is distributed as $N(\mu, \sigma^2/n)$, then the standardized Z-score:

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

Z has the std normal distribution with zero mean and unit std dev

- Standard normal distribution with 95% confidence level



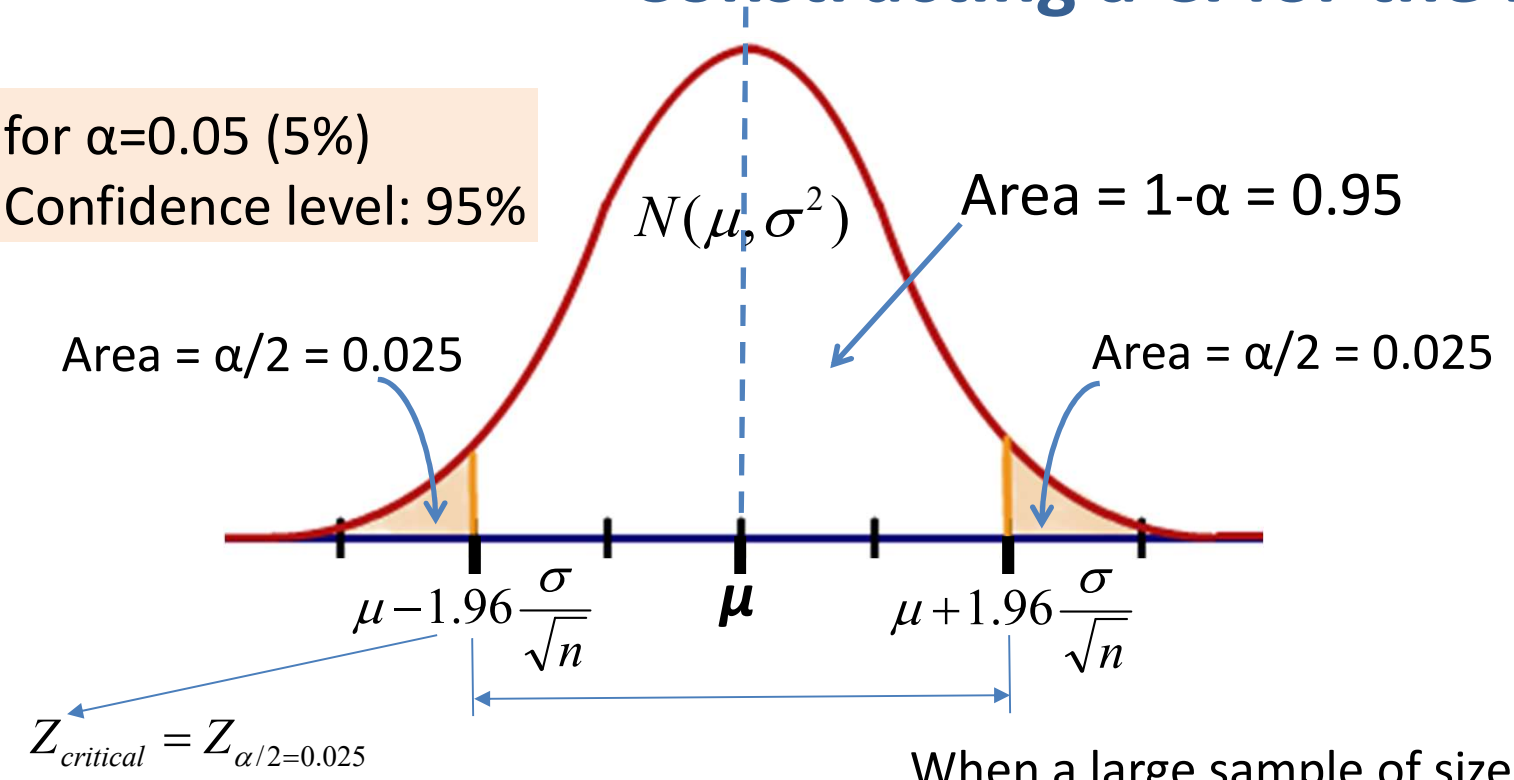
$$P(-1.96 \le Z \le 1.96) = 0.95 \ (area)$$

$$P(-1.96 \le \frac{\bar{X} - \mu}{\sigma / \sqrt{N}} \le 1.96) = 0.95$$

$$P(\mu - 1.96\frac{\sigma}{\sqrt{N}} \le \bar{X} \le \mu + 1.96\frac{\sigma}{\sqrt{N}}) = 0.95$$

Constructing a CI for the mean

for α=0.05 (5%)
Confidence level: 95%



$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 1.96\frac{\sigma}{\sqrt{n}}) = 0.95$$

When a large sample of size n is repeatedly taken, 95% of the time $\bar{X}$ will fall within $1.96\,\sigma/\sqrt{n}$ units of the population mean μ

$$P(\overbrace{\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}}^{\text{lower bound}} \le \mu \le \overbrace{\overline{X} + 1.96\frac{\sigma}{\sqrt{n}}}^{\text{upper bound}}) = 0.95$$

In the long run, 95 out of 100 CI's will contain the true population mean

Constructing a CI for the mean

- Constructing a 95% CI for the mean
  - Suppose we take a random sample size ***n*** from a population having a mean ***μ*** and variance ***σ²***. Then the interval is called a 95% CI for the mean assuming population variance σ^2 is known:

$$\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \;,\; \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

  - The number 1.96 in the formula is the z-score associated with an area of 0.475 under the standard normal curve. That is, 95% of the area falls between z_1=-1.96 and z_2=+1.96.
- Constructing a 90% CI for the mean
  - The z-score (from the table) that corresponds to an area of 0.45 is 1.645. Thus the formula for the 90% CI:

$$\left(\overline{X} - 1.645\frac{\sigma}{\sqrt{n}} \;,\; \overline{X} + 1.645\frac{\sigma}{\sqrt{n}}\right)$$

Constructing a CI for the mean

- Constructing a 99% CI for the mean
  - Similarly, the z-score that corresponds to an area of 0.495 is 2.58. Thus the formula for the 99% CI:

$$\left(\overline{X} - 2.58\frac{\sigma}{\sqrt{n}}\ ,\ \overline{X} + 2.58\frac{\sigma}{\sqrt{n}}\right)$$

- **Level of confidence of a CI**: If you draw a random sample many times, a certain percentage of the CI's will contain the population parameter. This percentage is known as the confidence level (denoted by 1-α).
- Confidence interval for a mean:

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Z-score such that the area to its right is α/2

Conf Level	$Z_{critical}$
90%	$Z_{\alpha/2=0.050} = 1.645$
95%	$Z_{\alpha/2=0.025} = 1.96$
99%	$Z_{\alpha/2=0.005} = 2.58$

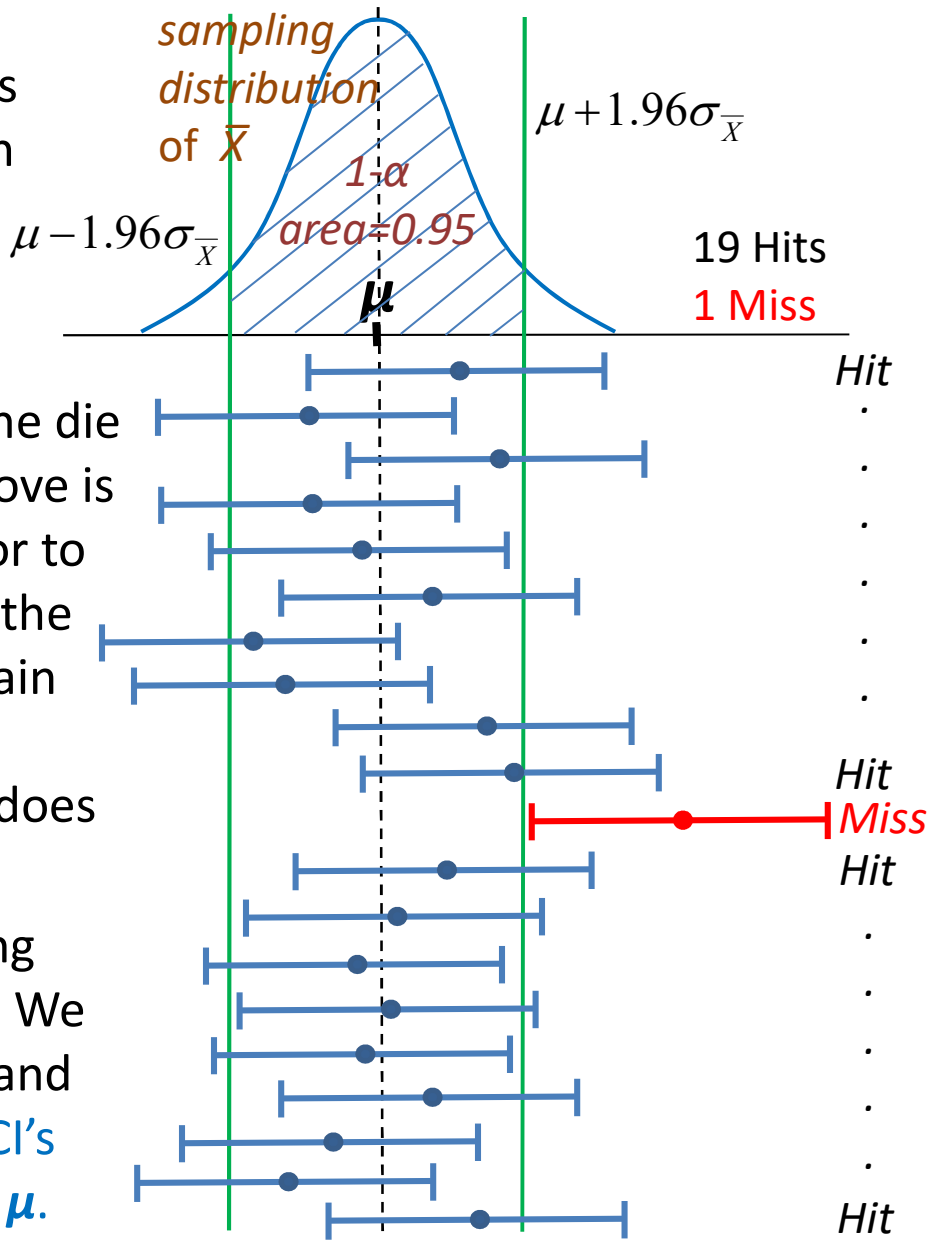
Interpretation of CI via simulation

A visual interpretation of the 95% CI for a population parameter $\boldsymbol{\mu}$ which is an unknown constant and remains an unknown constant (not a RV):

$$\overline{X} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}} = \overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Once $\overline{X}$ is actually observed, then "the die is cast", and the interval estimate above is either dead right or dead wrong. Prior to selecting a random sample of size ***n***, the probability is 95% that the CI we obtain will contain μ. After we've taken the sample, CI we've constructed either does or doesn't contain $\boldsymbol{\mu}$.

Repeat the entire process, keep taking samples of size ***n*** for, say 1000 times. We would have different sample means and CI's. We would expect 95% of these CI's to include the population parameter $\boldsymbol{\mu}$.



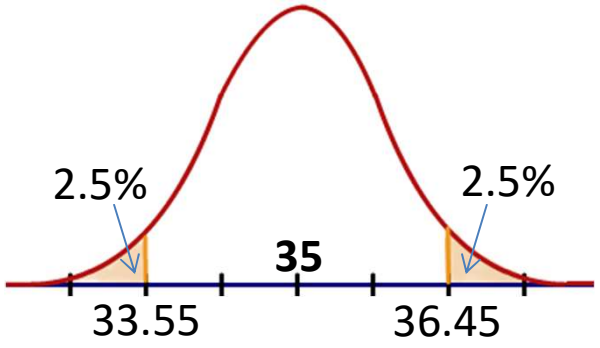
Example

- We want to estimate the mean amount spent on food per customer in a burger house. Data is collected for 90 customers. Assuming the population σ = 7 TL
  - What is the margin of error at 95% confidence?

$$\overline{X} \pm 1.96\,\sigma_{\overline{X}} = \overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \overline{X} \pm 1.96\frac{7}{\sqrt{90}} = \overline{X} \pm 1.45\,TL$$

  - If the sample mean is 35 TL, what is the confidence interval for the population mean (all customers)?

$$\overline{X} \pm 1.45\,TL = 35 \pm 1.45\,TL \;\Rightarrow\; [33.55\;,\;36.45]\;TL$$



All samples of size n = 90 will have 1.45 as the margin of error, assuming the pop std dev is 7

If we were to take 100 samples (of size 90), 95 of all CI's constructed using $\overline{X} \pm 1.45$ would contain the true population mean (unknown)

```
>>> conf_int = stats.norm.interval(0.95, 35, 0.7379)
```

SE = $7 / \sqrt{90}$

Confidence Intervals for proportions

- Standard Error for the sample proportion?
- A Binomial distribution B(n,p) is just the sum of Bernoulli trials with a success probability of p. If n is large enough (both np and n(1-p) ≥10), the CLT applies and you can approximate B(n,p) with a Normal distribution. If we have a Bernoulli random variable Y, such that:

$$\mathbf{Y} = \begin{cases} p & \text{success} \\ 1\text{-}p & \text{failure} \end{cases} \quad \text{with} \quad \mu_Y = p, \quad \sigma_Y{}^2 = p(1-p)$$

- If X is the sum of n indep. Bernoulli trials of Y, the expected value and the variance are $\mu_X = np$ and $\sigma_X{}^2 = np(1-p)$
- The expected value of our sampling distribution of our sample proportion p̂ (for np ≥10):

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

- Variance of p̂:

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}[np(1-p)] = \frac{p(1-p)}{n}$$

- Standard error for the sample proportion:

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Intervals for proportions

- To estimate the proportion of a population possessing a certain characteristic, we take a sample of size ***n*** and count the number of individuals (***X***) in the sample that possess the characteristic.
- Estimate of p (population proportion): $\hat{p} = X / n$
- **p̂** : A random variable that changes from sample to sample
- **p̂** : Approximately a normal distribution (given np ≥ 5, nq ≥ 5) with mean ***p*** and standard dev $\sqrt{pq/n}$, (SE) where ***q=1-p***.
- A 95% confidence interval is given by: $\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$ See previous slide
- **Example**: In a sample of 400 voters, 60% favored candidate A. What is a 99% confidence interval for the percentage of voters who favor candidate A?

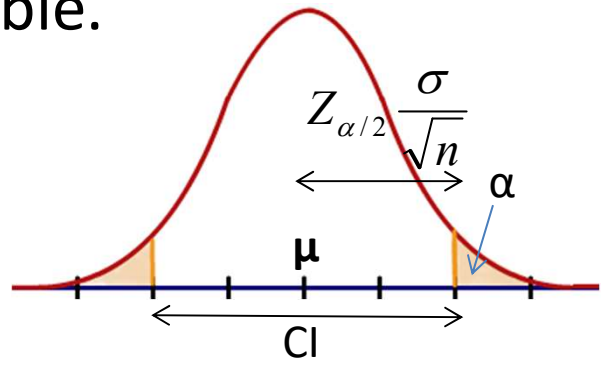
$$\left(0.6 - 2.58\sqrt{\frac{(.6)*(.4)}{400}}\ ,\ 0.6 + 2.58\sqrt{\frac{(.6)*(.4)}{400}} \right) = (0.537,\ 0.663)$$

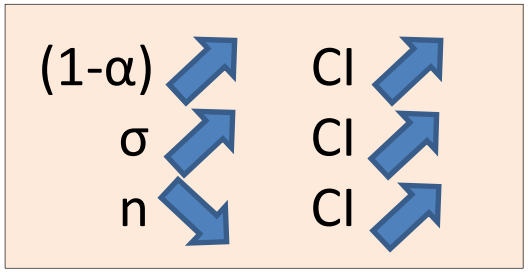
Level of confidence vs width of CI

- Tradeoff between the level of confidence and width of Confidence Interval
  - There is a tradeoff between the value of $\boldsymbol{\alpha}$ and the width of CI for a given sample size $\boldsymbol{n}$ and $\boldsymbol{\sigma^2}$.
  - The smaller the value of α (i.e., the larger the level of confidence 1- α), the smaller the tails of the standard normal curve must become (with larger values of $Z_{\alpha/2}$).
  - In most cases, we would like the confidence that our CI contains the true mean to be high, say 90% or more. We would also like our CI to be narrow so that the estimate is very precise.
  - One way of making the CI (for any value of α) as narrow as desired is to increase the sample size. Note that this will increase the cost of sampling too.
  - For a fixed sample size, the only way to increase the level of confidence is to increase the width of the CI as well.

Factors that influence the width of CI

- **Standard deviation (σ)**
  - CI will be wider when the underlying population has a larger standard deviation because more variability makes sample statistics less reliable.
- **Sample Size (n)**
  - The smaller the sample size you use, the wider the CI you get. So, you can estimate a population mean or proportion more accurately if you have a larger sample size.





- **Level of confidence (1- α)**
  - The greater the probability, the wider your CI will be. Thus a 99% CI is wider than a 95% CI. The more certain you want to be that the proportion or mean actually lies within the interval, the wider the interval must be.

Ideal sample size

- Ideal sample size to achieve a certain margin of error?
- Remember a CI takes the form: $\overline{X} \pm Z_{\alpha/2}\sigma/\sqrt{n}$
- Suppose we want $\bar{X}$ to lie within D units of μ with a level of confidence (1-α), where $D = Z_{\alpha/2}\sigma/\sqrt{n}$

- So the sample size (if mean) : $n = Z^2{}_{\alpha/2}\dfrac{\sigma^2}{D^2}$

- Sample size (if proportion) : $n = Z^2{}_{\alpha/2}\dfrac{\hat{p}\hat{q}}{D^2}$

- **Example**: Wish to estimate the mean of a population with a standard deviation of 5. To be 95% confident on our inference on the population mean with a margin of error ±1, we would need a sample of 96 observations (to have 95% of our sample means contain μ).
- **Example**: Wish to estimate the proportion of a variable that has a 40% observation rate in the population. To be 95% confident about our inference on population rate with a margin of error ±0.05, we need a sample of 369 observations.

CI for μ_1-μ_2

- **Confidence interval for population differences**
- The CI for the difference in means provides an estimate of the absolute difference in means of the outcome variable of interest between the comparison groups.
- Methodology:
  - Identify a sample statistic (like mean)
  - Use the difference between sample means to estimate the difference between population means
  - Select a confidence level (1-α)
  - Find the margin of error and specify the confidence interval
- Assumptions
  - Known population variances
  - Normally distributed populations
  - Independent random samples

CI for μ_1-μ_2 – cont'd

- We're interested in constructing CIs for the difference between 2 population means μ_1- μ_2
- Populations are normally distributed, so are the random variables X_1 and X_2. The difference between sample means $\bar{X}_1$ - $\bar{X}_2$ is also normally distributed with a **mean of (μ_1- μ_2)** and **variance of**: $\mathbf{Var}(\overline{X}_1 - \overline{X}_2) = \sigma^2_{\overline{X}_1 - \overline{X}_2}$

$$\mathbf{Var}(\overline{X}_1 - \overline{X}_2) = \mathbf{Var}(\overline{X}_1 + (-\overline{X}_2)) = \mathbf{Var}(\overline{X}_1) + (-1)^2\,\mathbf{Var}(\overline{X}_2)$$
$$= \sigma^2_{\overline{X}_1 - \overline{X}_2} = \sigma_{\overline{X}_1}{}^2 + \sigma_{\overline{X}_2}{}^2 = (\sigma_1{}^2 / n_1) + (\sigma_2{}^2 / n_2)$$

- A 95% CI for μ_1- μ_2 : $(\overline{X}_1 - \overline{X}_2) \pm 1.96\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}$
- **Important note**: So far, the calculation of the CI (for μ or μ_1- μ_2) is based on the assumption that the population variances are known. When the variances are unknown and have been estimated, then the calculation for the CI will be based on the t-distribution (coming soon)

CI for μ_1-μ_2 – cont'd

- **Example**: A mechanical device built by a company throws golf balls. Company claims that the device throws Type 1 balls farther. The device hits a sample of 225 golf balls of Type1 and 400 balls of Type2. The average distances the balls traveled are measured: $\bar{X}_1$ =204 m. and $\bar{X}_2$ =190 m. If we know that σ_1^2=450 and σ_2^2=800, what is the 95% confidence interval for μ_1- μ_2?

$$(\overline{X}_1 - \overline{X}_2) \pm 1.96\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}} = (204-190) \pm 1.96\sqrt{\frac{450}{225} + \frac{800}{400}}$$

$$= 14 \pm 3.92 = [17.92\ ,\ 10.08\ ]$$

- This experiment, before it was executed, had a 95% chance of producing an interval containing the true value. This result seems to be consistent with the assertion that Type1 balls travel farther than Type2 balls as the interval doesn't cross "0" point. If the CI includes "0", this implies that "0" is a reasonable possibility for the true value of difference.

CI for p_1-p_2

- **Confidence interval for difference in population proportions**
- To construct a confidence interval for p1-p2, the difference between 2 population proportions, we utilize the fact that the estimator $\hat{p}_1 - \hat{p}_2$ is approximately normally distributed with

Mean : $p_1 - p_2$

Variance: $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

- We again utilize the standard normal variable Z when we construct the confidence intervals and obtain the following:
- A 95% CI for $\hat{p}_1 - \hat{p}_2$: $(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

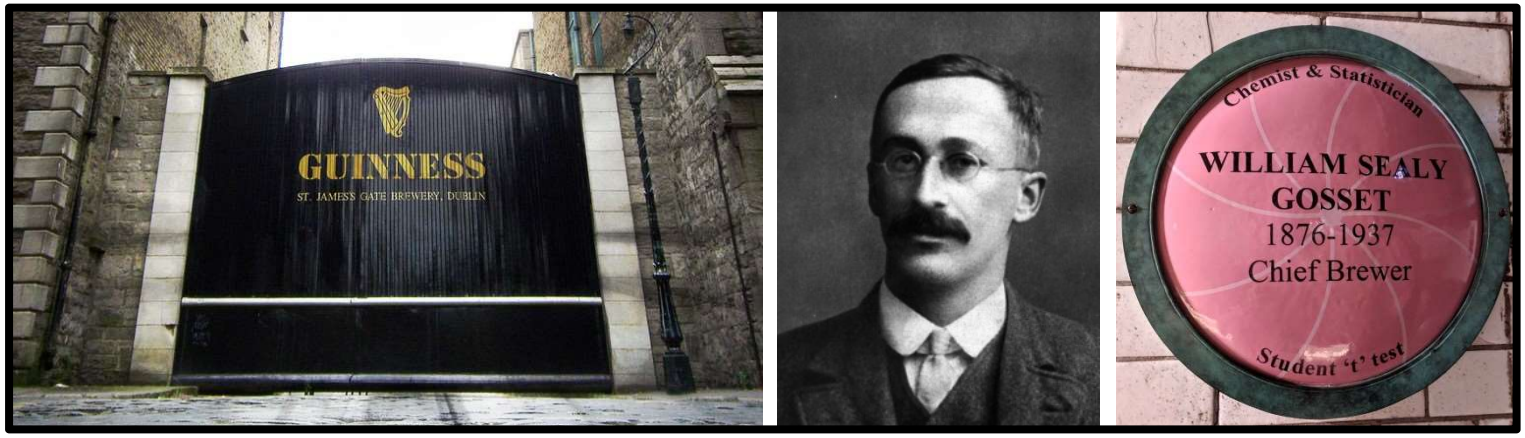
t-distribution and Confidence Intervals

- Until now, we discussed that sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and variance σ^2/n
  - The population is normally distributed
  - The sample size is large, say n ≥ 30 ($\bar{X}$ has approximately normal distribution and the approximation improves as n increases by the CLT)
  - If $\bar{X}$ is distributed as $N(\mu, \sigma^2/n)$, then the standardized z-score: $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ (a standard normal variable)
- If the variance is known, we can construct a CI for the mean using the z-score and the std normal distribution.
- The population variance (std dev), however, is usually unknown and is estimated by the sample variance S^2:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

t-distribution and Confidence Intervals

- Z-test (normal distribution) works great for relatively large sample sizes. What if the sample size is really small, say, just a few?
- William S. Gosset, the chief brewer of Guinness, had a problem with small sample sizes that he used to figure out the sugar content in the malt extract.



- Gosset created a distribution to account for the sample size effect.

Student's t-distribution

VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

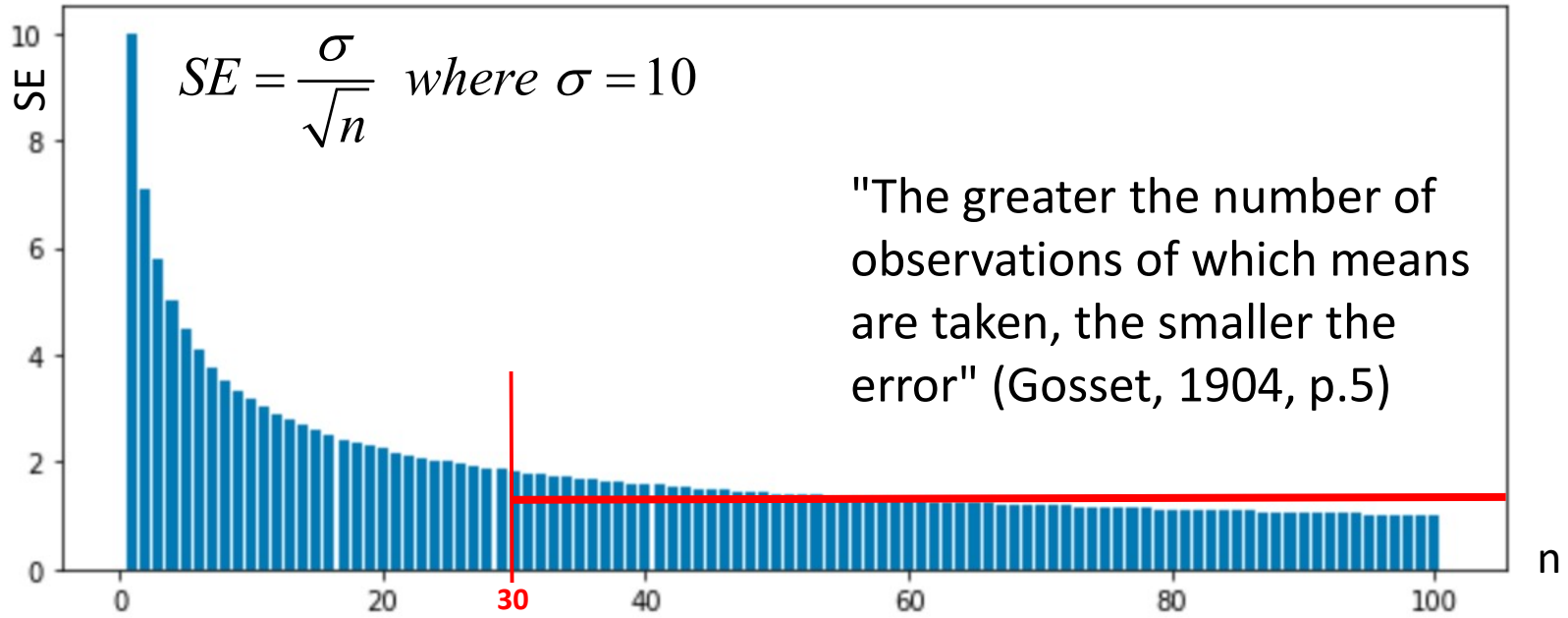
ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

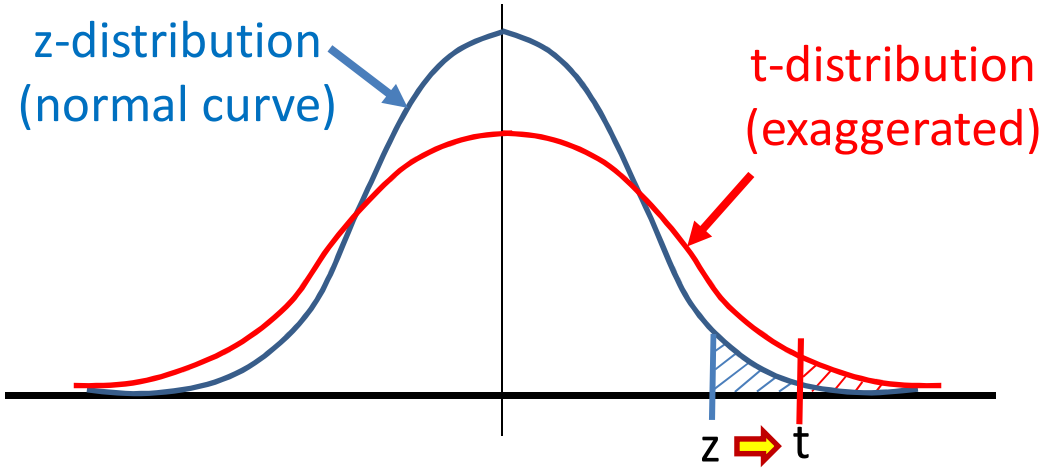
If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in

t-distribution and Confidence Intervals

- Effect of sample size on Standard Error



"The greater the number of observations of which means are taken, the smaller the error" (Gosset, 1904, p.5)



The critical value **z** is pushed further out to **t**, so fatter tails accommodate more extreme values to account for the uncertainty in the SE (for small sample sizes).

t-distribution and Confidence Intervals

- Replacing σ by S in the z-score, we obtain what is called the t-score:

$$t = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

- If the population is normally distributed, then the t-score follows the t-distribution (aka **Student's t-distribution**). The use of t-distribution is based on 2 assumptions:
  1. X is normally distributed
  2. S^2 has been used to estimate σ^2
- t-score then follows the t-distribution (at least approximately) if any of the following 3 conditions apply:
  1. Population is normal and sample is large (obvious)
  2. Population is normal and sample is small
  3. Population is non-normal and sample is large (this result is only an approximation)

t-distribution (Student's t-distribution)

- Characteristics of t-distribution
  - Has a mean of 0
  - Distribution is symmetrical about the mean 0 (bell-shaped)
  - Distribution depends on a parameter called **degrees of freedom** (d.f.) denoted by ***v***

  d.f. = ***v*** = n – 1 (n: number of observations in the sample)
- Student's t-distribution has the pdf given by:

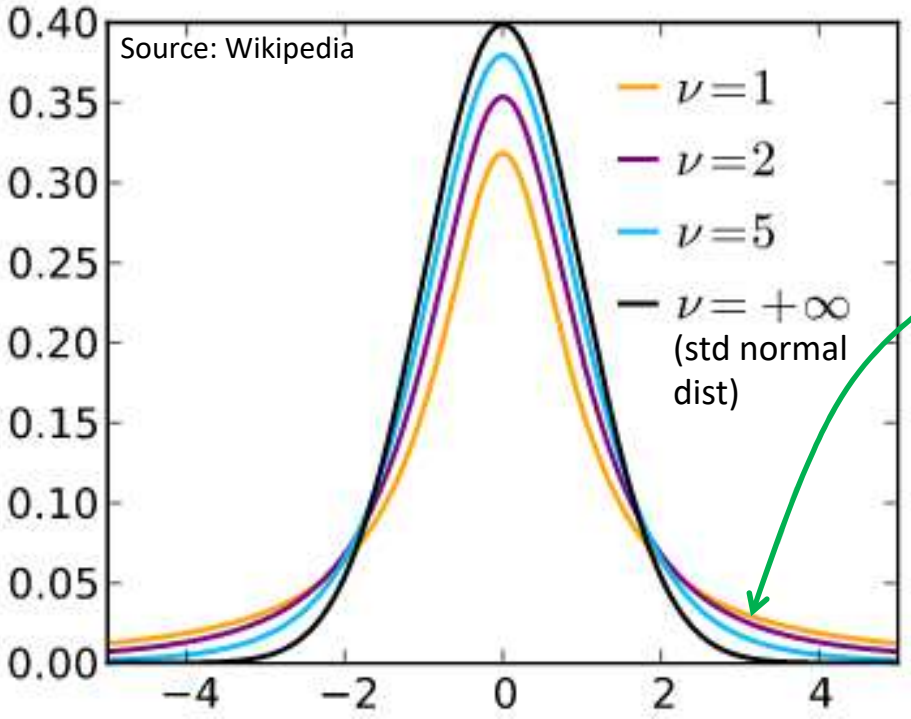
$$f(t)=\frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\,\Gamma\left(\frac{\upsilon}{2}\right)}\left(1+\frac{t^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}} \qquad -\infty<t<\infty$$

where Γ is the Gamma function

  - When population **σ** is unknown, it's replaced with s. Computation of **s** requires computation of the sample mean which consumes **1** degree of freedom leading to **df = n-1**.

t-distribution (Student's t-distribution)

- The t-distribution deals with the extra unknown **σ** which is estimated by the random variable S. So, it tacitly handles greater uncertainty in the data. As a consequence, the t distribution has wider confidence intervals than the normal (z) distribution.



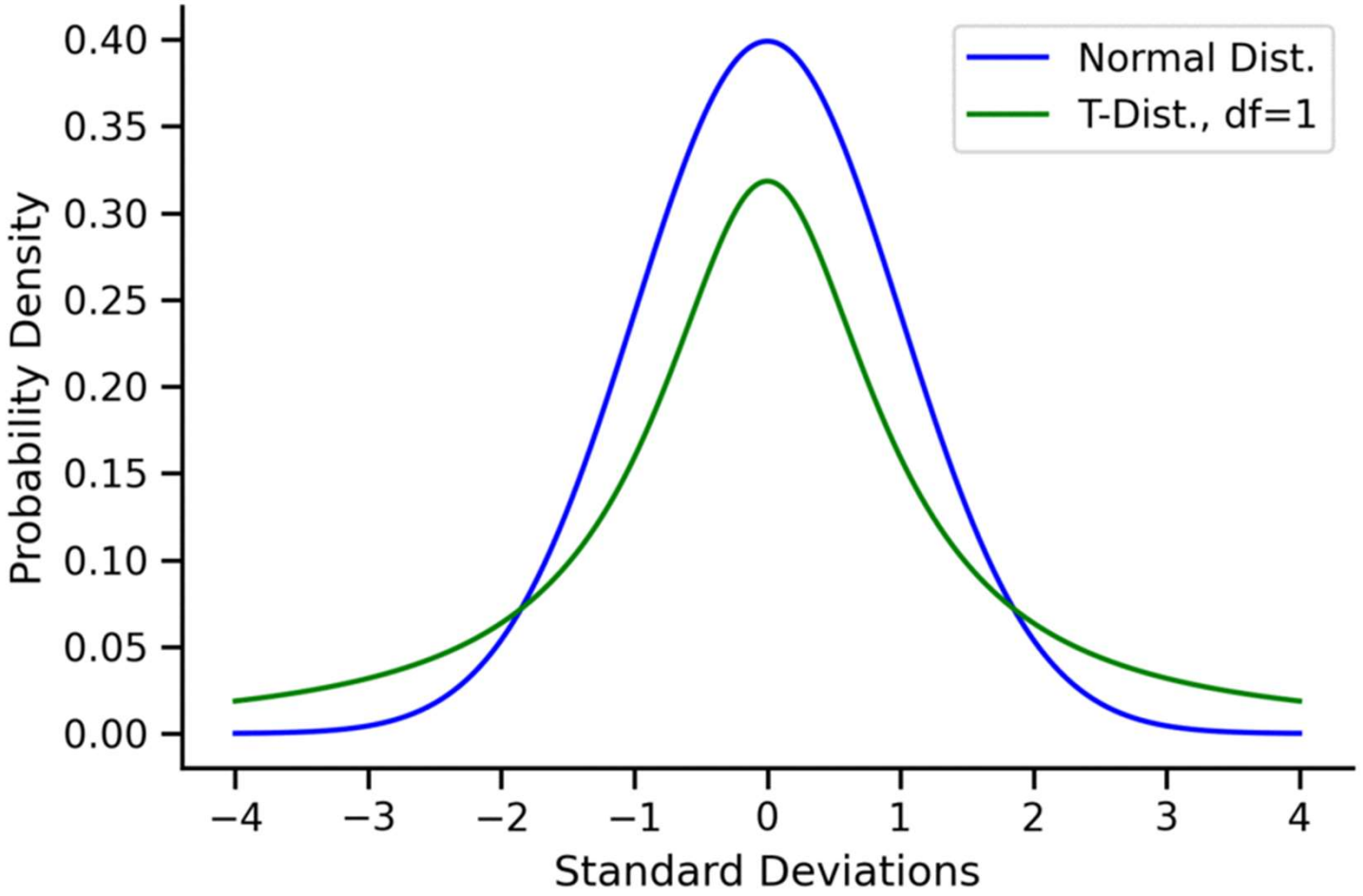
As the degrees of freedom (***v***) increases, t-distribution converges to normal distribution

t-distribution is used for smaller sample sizes to account for the uncertainty in the standard dev. of the sample. Fat tails indicate that probability of observing extreme values are higher.

For interactive visualization: http://rpsychologist.com/d3/tdist/

t-distribution (Student's t-distribution)

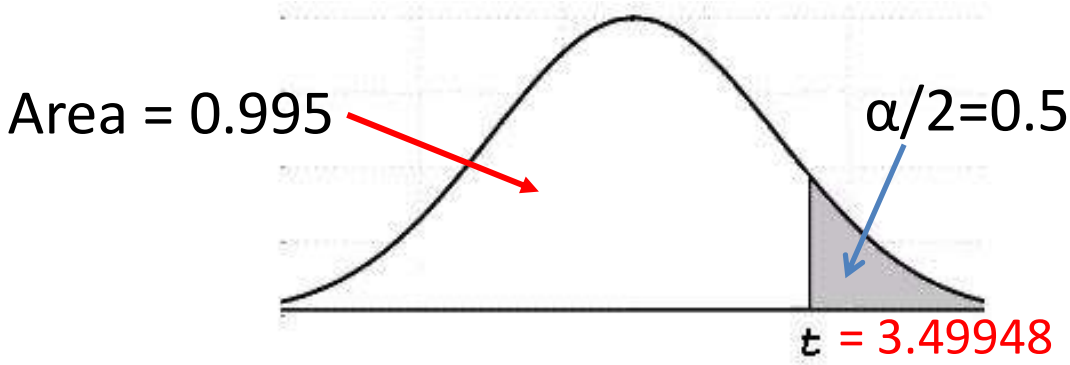
Source: https://tjkyner.medium.com/the-normal-distribution-vs-students-t-distribution-322aa12ffd15



t-distribution doesn't need any input regarding the population mean and standard deviation, which are seldom known anyway, and it has the ability to account for the uncertainty due to small sample sizes which makes it attractive for most cases.

How to read the t-value from a t-table?

- t-value for α=0.01 (99% confidence) and df=7 ?



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413

- Python code (ppf of the t-dist):

```
>>> print(stats.t.ppf(0.995, 7))
3.49948329735
```

Percent point function at 99.5% with df=7

Confidence intervals for μ

Confidence level	Interval
90%	$\overline{X} \pm t_{0.05} S / \sqrt{n}$
95%	$\overline{X} \pm t_{0.025} S / \sqrt{n}$
99%	$\overline{X} \pm t_{0.005} S / \sqrt{n}$

- **Example**: On 8 Monday mornings, a count is made of the number of cars crossing a bridge between 7:00 and 9:00 am with $\overline{X}$ = 1500 and S^2=80000. Assuming the population is normal, construct a 99% confidence interval for the population mean.
  - Solution: There are n - 1 = 7 degrees of freedom and $t_{0.005}$ = 3.5. So a 99% confidence interval is given by:

$$\overline{X} \pm t_{0.005} S / \sqrt{n} = 1500 \pm (3.5) * (100) = (1150 \ , 1850)$$

Confidence Interval for μ_1-μ_2

- **Case 1:** with **known variances** (σ_1^2,σ_2^2):

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$$

- **Case 2:** with **unknown but equal variances** ($\sigma_1^2 = \sigma_2^2$):

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{S_p^{\,2}}{n_1} + \frac{S_p^{\,2}}{n_2}}$$ with dof = $n_1 + n_2 - 2$

  - where S_p^2 is the pooled variance estimate (a weighted average of sample variances) given as:

$$S_P^{\,2} = \frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{(n_1 - 1) + (n_2 - 1)}$$

  - and $t_{\alpha/2}$ is the t-value having area $\alpha/2$ to its right. The appropriate number of degrees of freedom is n_1+n_2-2

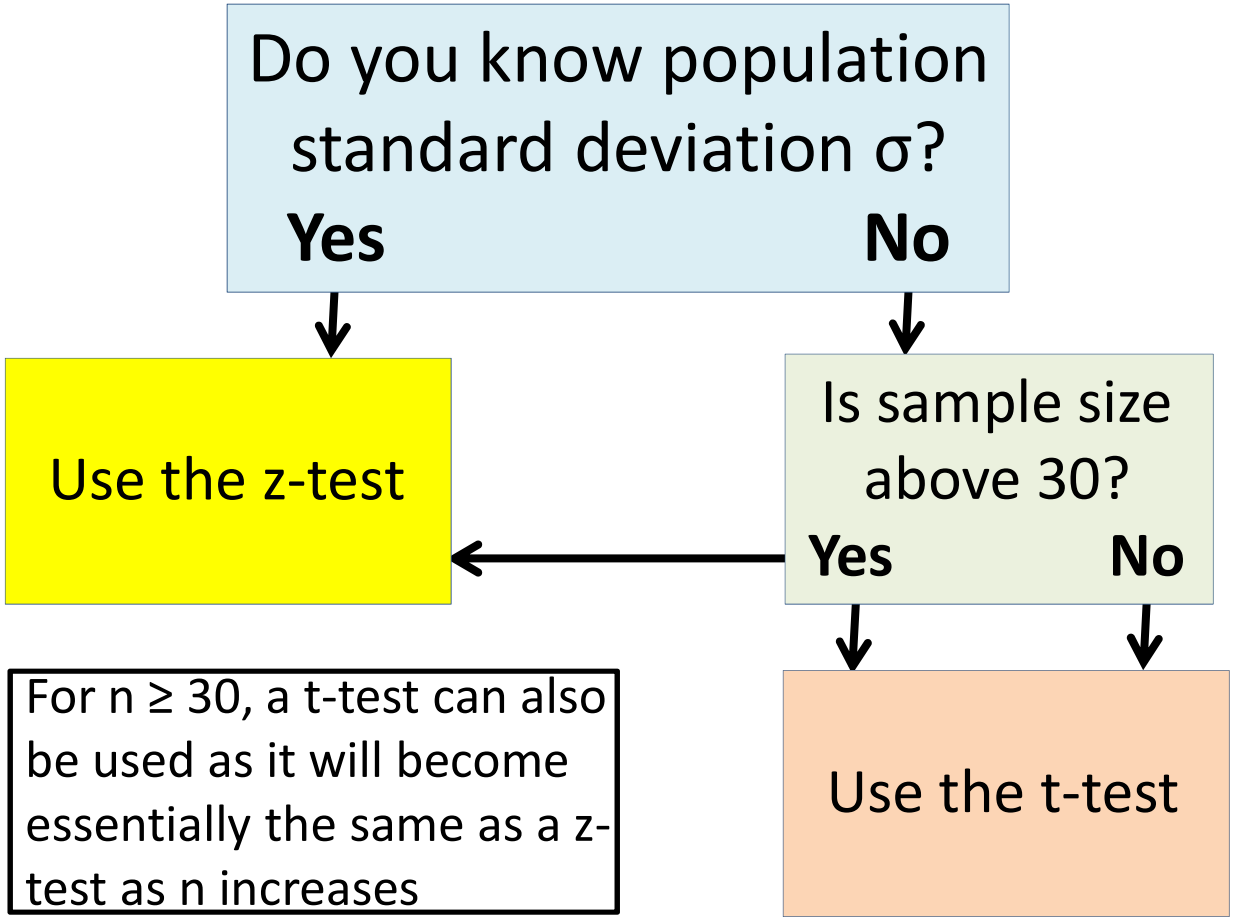
Confidence Interval for μ_1-μ_2

- **Case 3:** with **unknown & unequal variances** ($\sigma_1^2 \neq \sigma_2^2$):

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad \text{with} \quad df = \frac{[(S_1^2/n_1) + (S_2^2/n_2)]^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

- When to use equal or unequal variances?
- Short answer: Use unequal variances (conservative choice)
- Use equal variance when
  - You know the population variances are equal based on prior knowledge/analysis, or you have a good reason to believe so
- Use unequal variances when
  - You know the population variances are not the same
  - You DON'T know if the variances are the same or not

When to use z-score and t-score?



The t-test assumes the population is normally distributed. However, it is fairly robust to violations of this assumption for sample sizes equal to or greater than 30, provided the observations are collected randomly and the data are continuous, unimodal, and reasonably symmetric.

Example I

- Is there a significant evidence that the average age of runners in 2019 has changed significantly with respect to the average age of all past events in the Big Charity run?
- Past : $\mu_{<2019} = 36.13$ years
- 2019 : $\mu_{2019} = 35.05$ years with s=8.97 for 100 runners
- Let's construct a confidence interval for 95%:

```
stats.t.ppf(0.975, 99)
>>> 1.984
```

$$\overline{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = \overline{X} \pm 1.984 \frac{8.97}{\sqrt{100}} = \overline{X} \pm 1.78$$

- $\overline{X} \pm 1.78 = 35.05 \pm 1.78$ => (33.27 , 36.83)
- This interval contains the average age for the past events and we can say that the change in age is **NOT significant**.

Example II

- Over a period of time, 200 patients came to a doctor complaining about severe headaches.
- Doctor gave **100 patients a special headache pill** and **35** of them **improved within a few hours**.
- Doctor gave the other **100 patients a placebo** and **30** of them **improved within a few hours** in this sample.
- Is there a significant difference between the two cases?
- Let's find a 99% confidence interval for $\hat{p}_1 - \hat{p}_2$, the difference between the proportions of 2 groups of patients.
- Proportions for two groups: $\hat{p}_1 = 0.35$ and $\hat{p}_2 = 0.30$
- Construct a 99% CI for $\hat{p}_1 - \hat{p}_2$ ($Z_{\alpha/2}=Z_{0.005}=2.58$ for $1-\alpha=0.99$):

$$(0.35-0.30) \pm 2.58\sqrt{\frac{0.35*0.65}{100}+\frac{0.30*0.70}{100}}$$

$$= 0.05 \pm 0.17 = [-0.12\ ,\ 0.22]$$

Since the interval contains the zero value, we cannot conclude that the pill is more effective than the placebo.

Example III

- Different books are used to teach Statistics
- 20 of randomly selected students used Book1
- 15 of randomly selected students used Book2
- Students are approximately equal in ability and the average test scores are $\bar{X}_1 = 78$ and $\bar{X}_2 = 84$. Variances are assumed to be equal and were estimated to be $S_1^2 = 41$ and $S_2^2 = 36$.
- Construct a 95% CI for $\mu_1 - \mu_2$:

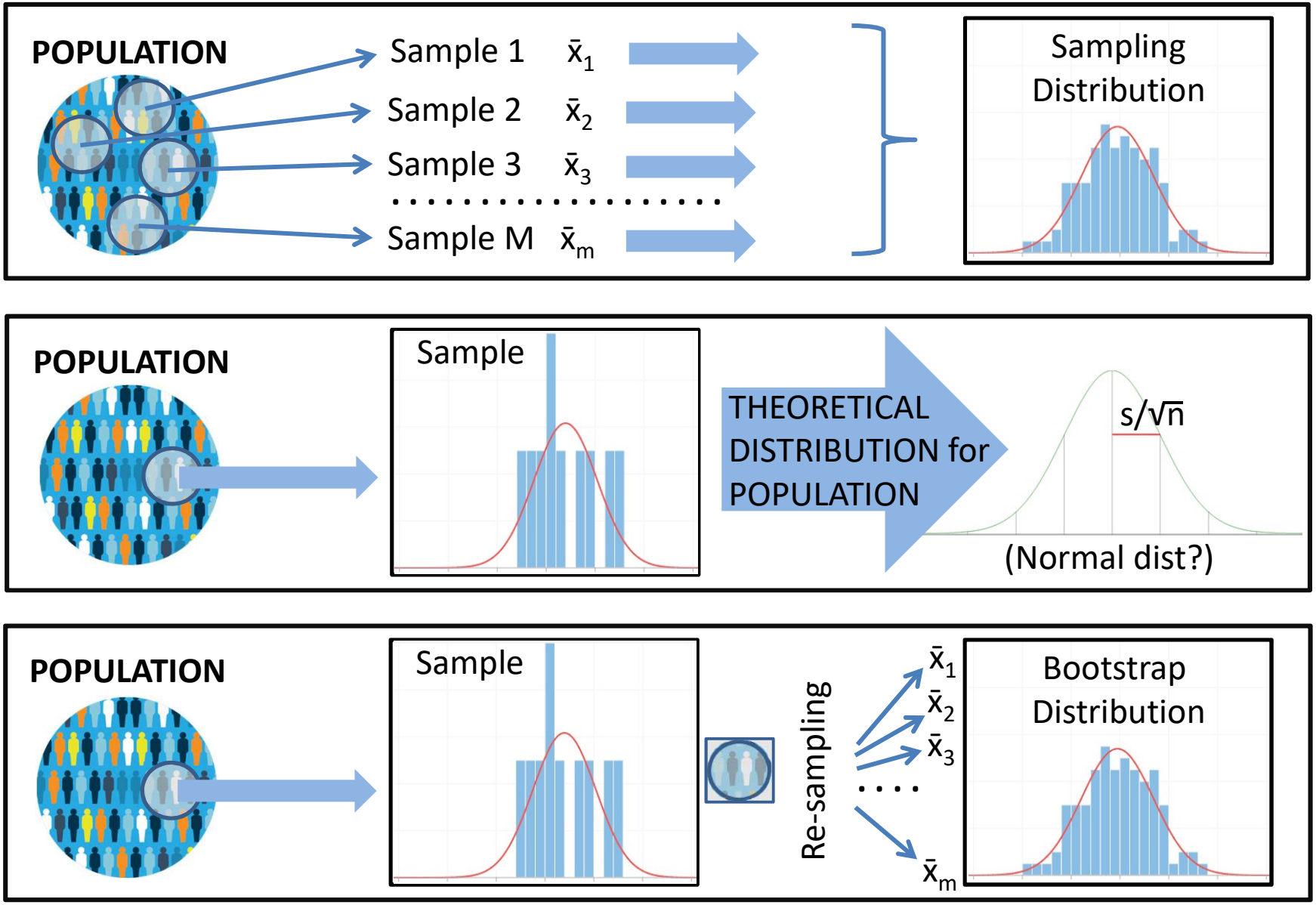
$$S_P^{\,2} = \frac{(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}}{(n_1 - 1) + (n_2 - 1)} = \frac{(20 - 1)41 + (15 - 1)36}{(20 - 1) + (15 - 1)} = 38.88$$

We have the level of confidence $1-\alpha=0.95$. For (20+15-2) = 33 degrees of freedom, we obtain approximately $t_{\alpha/2} = 2.04$. The 95% CI for $\mu_1-\mu_2$ is given by:

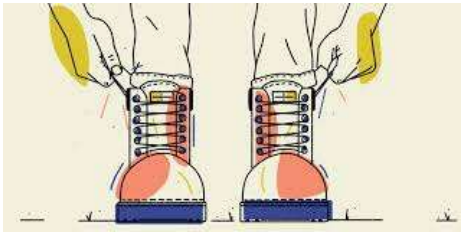
interpretation?

$$(78 - 84) \pm 2.04\sqrt{\frac{38.88}{20} + \frac{38.88}{15}} = (-6 \pm 4.345) \Rightarrow (-10.345, -1.655)$$

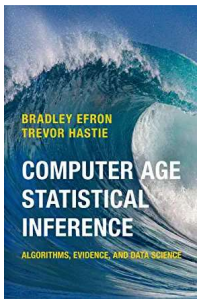
Bootstrapping

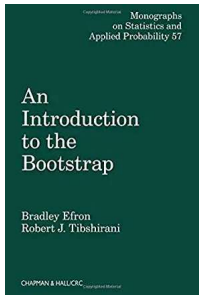


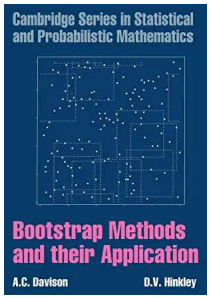
Bootstrapping

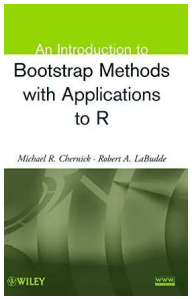


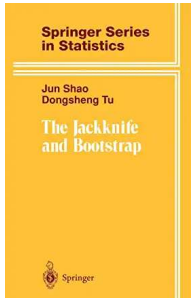
- Comes from the term "pulling yourself up by your own bootstraps."
- Known as a self-starting process without any external input (resides in a computer ROM)
- In statistics, however, bootstrapping refers to **a technique used to estimate a population parameter by randomly resampling from the same sample with replacement** (building a system using itself).
- Bootstrapping is well-studied and rests on solid grounds
- Works poorly with very few samples (n > 30)

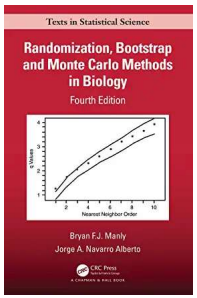


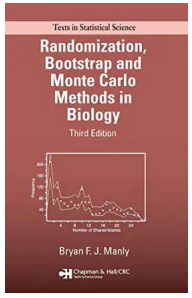






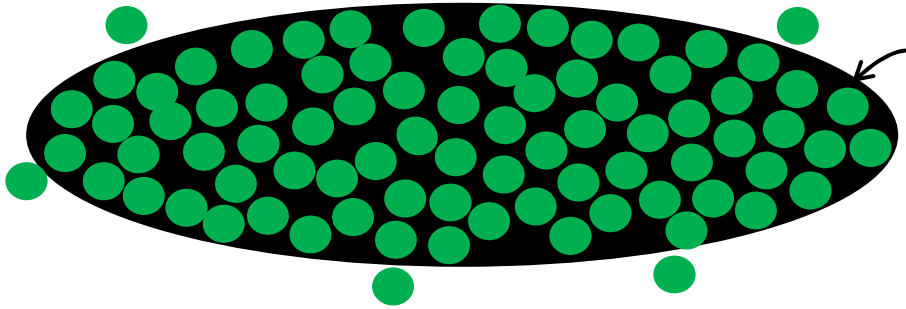




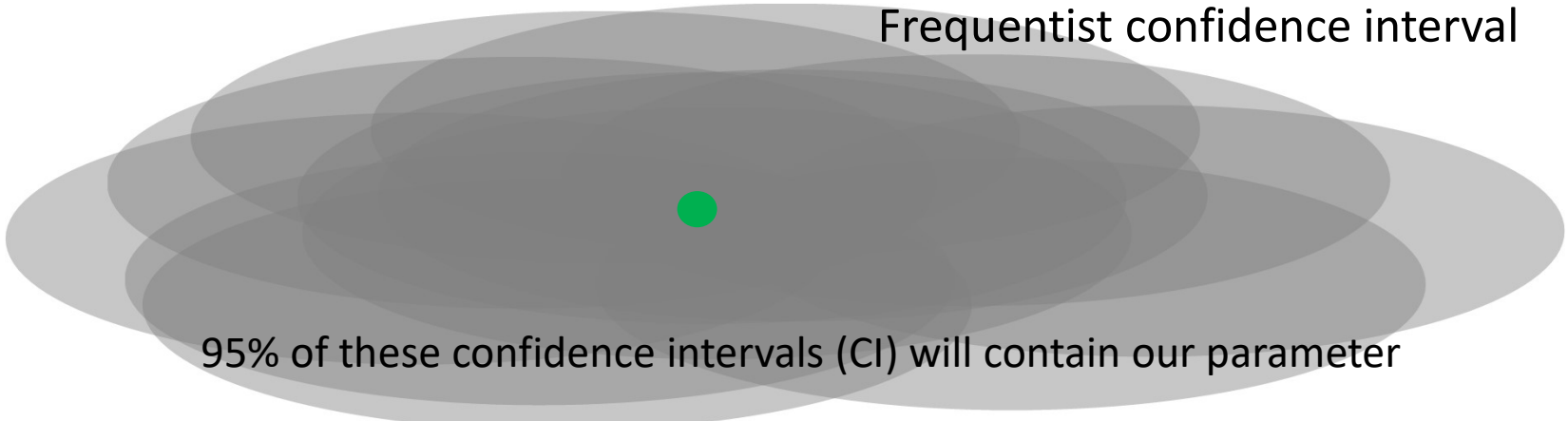


Confidence interval from two perspectives

- **Bayesian approach**: Probabilistic statement about model parameters given a fixed credible region



- **Frequentist approach**: Probabilistic statement about a recipe for generating confidence intervals given a fixed model parameter



From: Frequentism and Bayesianism: What's the Big Deal? | SciPy 2014 | Jake VanderPlas