

Confidence Intervals

Parthiban Rajendran

October 12, 2018

Contents

1	Theory	2
1.1	CI Formula Evolution	2
1.1.1	Example 1: Uniform Random Variable with 100% CI	2
1.1.2	Example 1b: Uniform Random Variable with 90% CI	3
1.1.3	Example 2: Uniform Random Variable with 100% CI	5
1.1.4	Example 3: Normal Distribution with 95% CI	7
1.2	CI for Sampling Distribution	10
1.2.1	95% CI as a Corollary	10
1.2.2	Generalized CI	11
1.2.3	When σ is known	12
1.2.4	When σ is not known	13
1.2.5	CI for difference between two means	13
1.2.6	CI for difference between two proportions	16
2	Examples	17
2.1	Deep Examples	17
2.1.1	Confidence Intervals for Sampling Proportions	17
2.1.2	Confidence Intervals for Sample Means	24
2.2	Shallow Examples	31
2.2.1	σ Known, Population Normal, Low Sample Size	31
2.2.2	σ Known, Population not Normal, High Sample Size	32
2.2.3	σ Unknown, Population Normal, Low Sample Size	33
2.2.4	σ Unknown, Population not Normal, High Sample Size	34
2.2.5	Difference between two means, Welch's 't' interval	35
2.2.6	Difference between two proportions	37
2.3	Useful Snippets	38
2.3.1	Python	38
2.3.2	Tikz in Ipython	39
3	Appendix	43
3.0.1	Difference between two Random Variables	43

Chapter 1

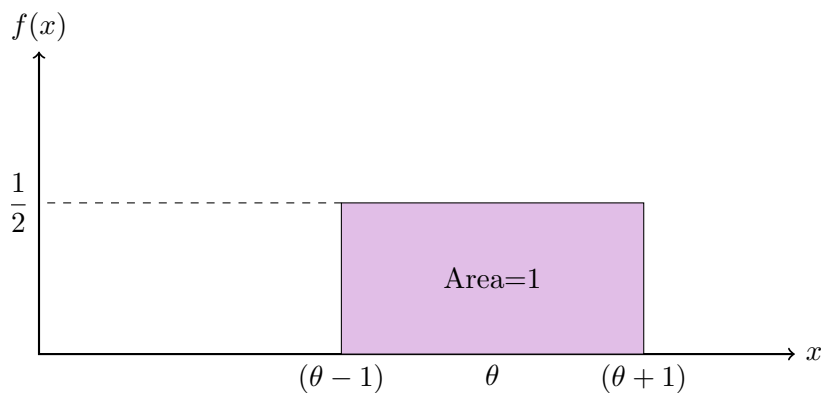
Theory

1.1 CI Formula Evolution

1.1.1 Example 1: Uniform Random Variable with 100% CI

Initial Setup

Random Variable x having uniform probability density function $f(x)$.



This simply means, the converge probability,

$$Pr(\theta - 1 \leq x \leq \theta + 1) = 1 \quad (1.1)$$

That is, the probability that x could be within $\theta \pm 1$ is 1.

CI construction using Pivotal Quantity

In equation 1.1, by adding $-\theta$ to the inequalities, we get,

$$\begin{aligned} Pr(-\theta + \theta - 1 \leq -\theta + x \leq -\theta + \theta + 1) &= 1 \\ Pr(-1 \leq x - \theta \leq 1) &= 1 \end{aligned} \quad (1.2)$$

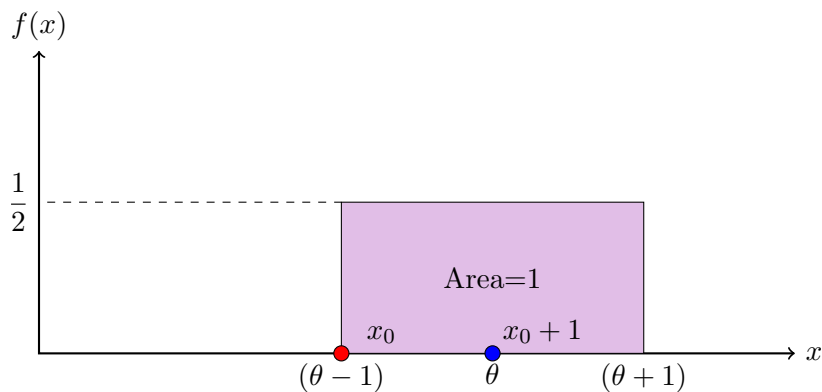
Multiplying by -1 , and adding x

$$\begin{aligned}
Pr(1 \geq -x + \theta \geq -1) &= 1 \\
Pr(x + 1 \geq \theta \geq x - 1) &= 1 \\
Pr(x - 1 \leq \theta \leq x + 1) &= 1
\end{aligned} \tag{1.3}$$

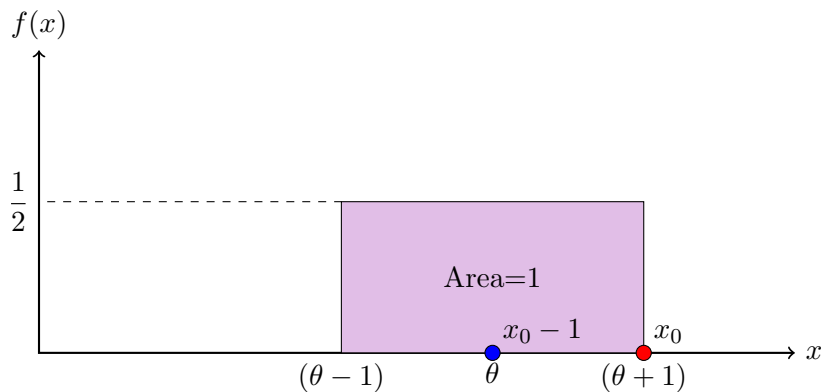
Thus while x could take value only between $\theta \pm 1$ for given probability, Equation 1.3 states, θ could also be only within $x \pm 1$ for same probability

Intuitive Proof

Suppose x takes a left extreme value as below within bounds $\theta \pm 1$.



Then, we could already see, θ is at $x_0 + 1$ still respecting the bounds $x \pm 1$. Suppose x takes a right extreme value as below within bounds $\theta \pm 1$.



Then, we could already see, θ is at $x_0 - 1$ still respecting the bounds $x \pm 1$.

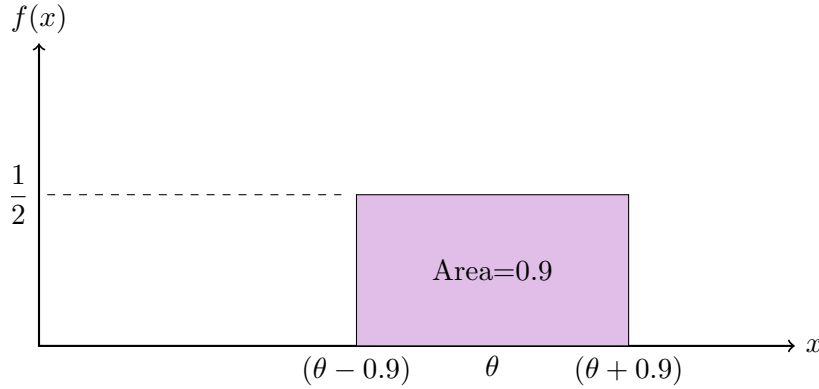
Thus, while x could be only within $\theta \pm 1$, it is also valid to say, θ could vary only within $x \pm 1$.

$$Pr(\theta - 1 \leq x \leq \theta + 1) = Pr(x - 1 \leq \theta \leq x + 1) = 1 \tag{1.4}$$

1.1.2 Example 1b: Uniform Random Variable with 90% CI

Initial Setup

Random Variable x having uniform probability density function $f(x)$.



This simply means, the converge probability,

$$Pr(\theta - 0.9 \leq x \leq \theta + 0.9) = 0.9 \tag{1.5}$$

That is, the probability that x could be within $\theta \pm 0.9$ is 0.9 or 90%.

CI construction using Pivotal Quantity

In equation 5, by adding $-\theta$ to the inequalities, we get,

$$\begin{aligned} Pr(-\theta + \theta - 0.9 \leq -\theta + x \leq -\theta + \theta + 0.9) &= 0.9 \\ Pr(-0.9 \leq x - \theta \leq 0.9) &= 0.9 \end{aligned} \tag{1.6}$$

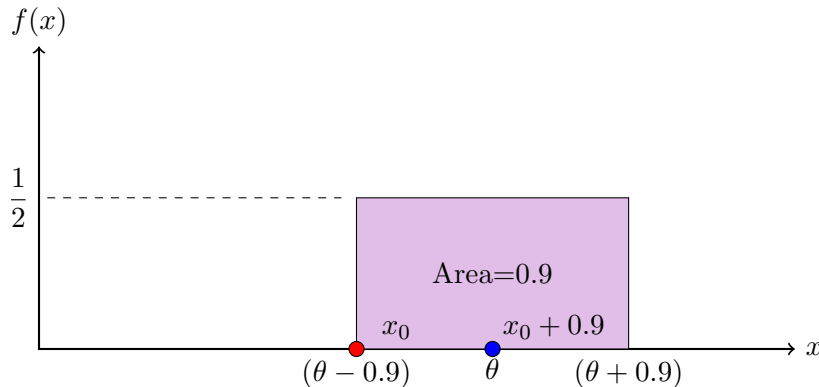
Multiplying by -1 , and adding x

$$\begin{aligned} Pr(0.9 \geq -x + \theta \geq -0.9) &= 0.9 \\ Pr(x + 0.9 \geq \theta \geq x - 0.9) &= 0.9 \\ Pr(x - 0.9 \leq \theta \leq x + 0.9) &= 0.9 \end{aligned} \tag{1.7}$$

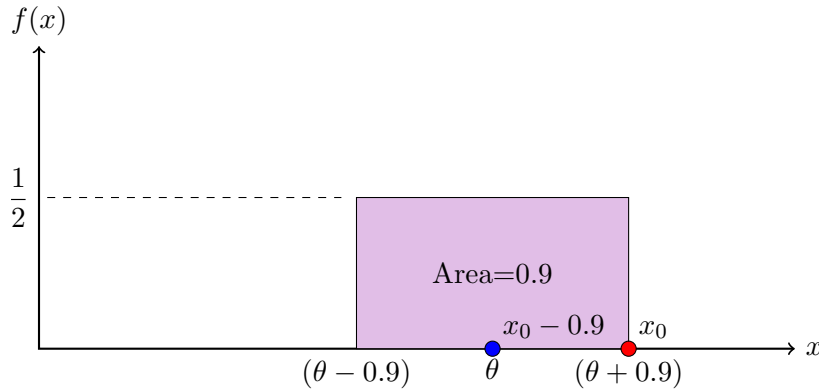
Thus, while x could take value only between $\theta \pm 0.9$ for given probability 0.9, above equation states, θ could also be only within $x \pm 0.9$ for same probability.

Intuitive Proof

Suppose x takes a left extreme value as below within bounds $\theta \pm 0.9$.



Then, we could already see, θ is at $x_0 + 0.9$ still respecting the bounds $x \pm 0.9$.
 Simply put, when x is at $x_0 = \theta - 0.9$, then it automatically implies, $\theta = x_0 + 0.9$
 Suppose x takes a right extreme value as below within bounds $\theta \pm 1$.



Then, we could already see, θ is at $x_0 - 0.9$ still respecting the bounds $x \pm 0.9$.

Thus, while x could be only within $\theta \pm 0.9$, it is also valid to say, θ could vary only within $x \pm 0.9$.

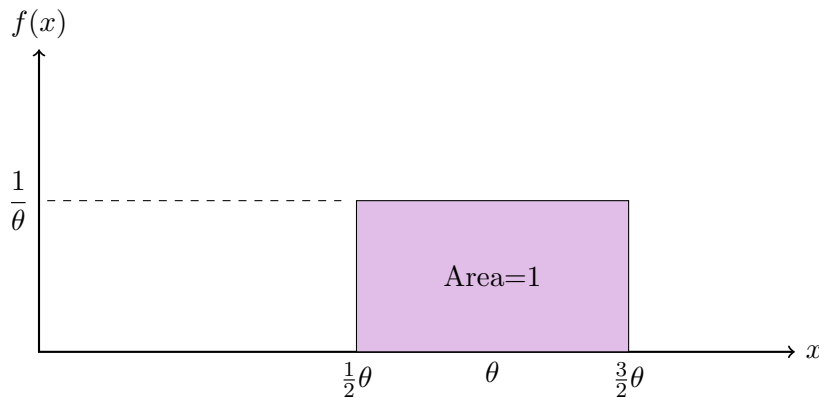
$$Pr(\theta - 0.9 \leq x \leq \theta + 0.9) = Pr(x - 0.9 \leq \theta \leq x + 0.9) = 0.9 \tag{1.8}$$

1.1.3 Example 2: Uniform Random Variable with 100% CI

Initial Setup

Random Variable x having uniform probability density function

$$f(x) = \frac{1}{\theta} \text{ for } \frac{1}{2}\theta \leq x \leq \frac{3}{2}\theta \tag{1.9}$$



This simply means, the converge probability,

$$Pr\left(\frac{1}{2}\theta \leq x \leq \frac{3}{2}\theta\right) = 1 \tag{1.10}$$

That is, the probability that x could be within $\theta \pm \frac{1}{2}\theta$ is 1

CI construction using Pivotal Quantity

Multiplying by 2 in the inequalities,

$$Pr(\theta \leq 2x \leq 3\theta) = 1$$

Dividing by θ ,..

$$Pr\left(1 \leq \frac{2x}{\theta} \leq 3\right) = 1$$

Dividing by x and inverting the inequalities, and again multiplying by 2..

$$Pr\left(\frac{1}{x} \leq \frac{2}{\theta} \leq \frac{3}{x}\right) = 1$$

$$Pr\left(x \geq \frac{\theta}{2} \geq \frac{x}{3}\right) = 1$$

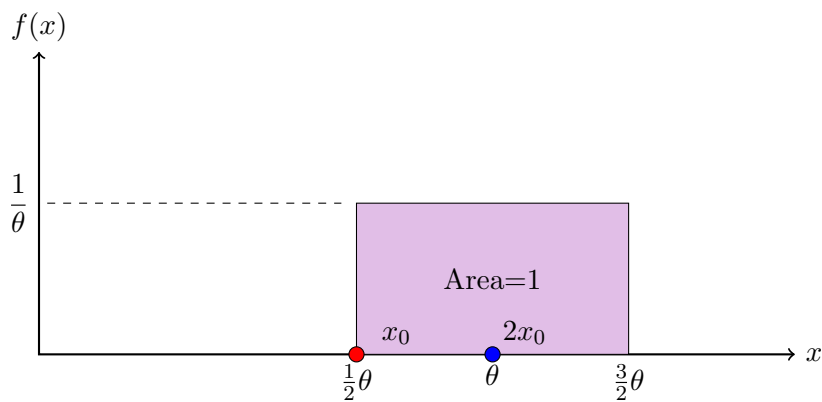
$$Pr\left(2x \geq \theta \geq \frac{2x}{3}\right) = 1$$

which is same as

$$Pr\left(\frac{2x}{3} \leq \theta \leq 2x\right) = 1 \tag{1.11}$$

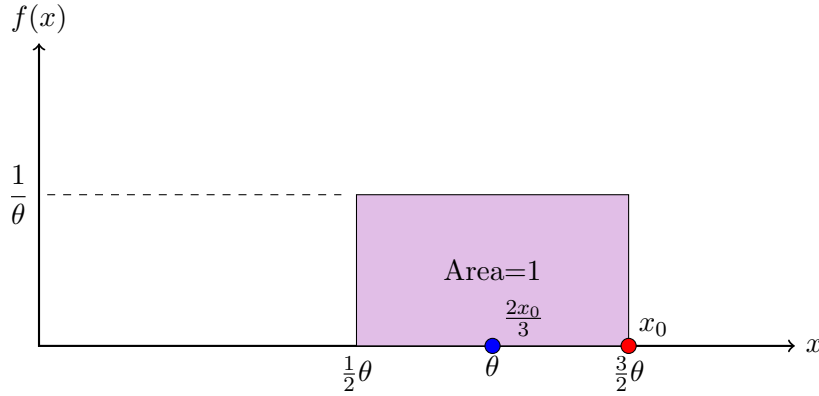
Intuitive Proof

Suppose x takes a left extreme value as below within bounds $\theta \pm \frac{\theta}{2}$.



When x is at $x_0 = \frac{\theta}{2}$, then $\theta = 2x_0$

Suppose x takes a right extreme value as below within bounds $\theta \pm \frac{\theta}{2}$.



When x is at $x_0 = \frac{3\theta}{2}$, then $\theta = \frac{2x_0}{3}$

Thus, while x could be only within $\theta \pm \frac{\theta}{2}$, it is also valid to say, θ could vary only within $(\frac{2x}{3}, 2x)$.

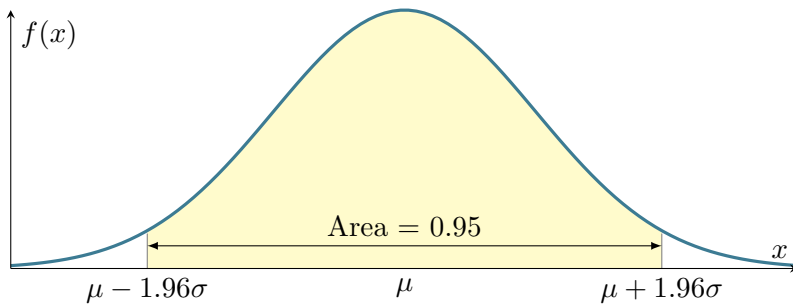
$$Pr\left(\theta - \frac{\theta}{2} \leq x \leq \theta + \frac{\theta}{2}\right) = Pr\left(\frac{2x}{3} \leq \theta \leq 2x\right) = 1 \tag{1.12}$$

1.1.4 Example 3: Normal Distribution with 95% CI

Initial Setup

Random Variable x having uniform probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{1.13}$$



This simply means, the converge probability,

$$Pr(\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma) = 0.95 \tag{1.14}$$

That is, the probability that x could be within $\mu \pm 1.96\sigma$ is 0.95 or 95%

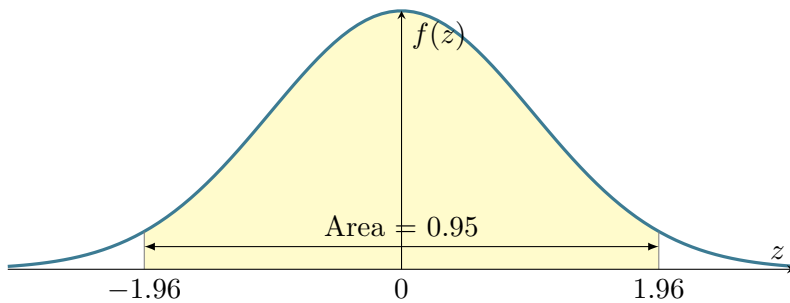
Why 1.96?

Let us standardize the distribution to standard normal distribution, $Z = \frac{X - \mu}{\sigma}$.

When $X = \mu + 1.96\sigma$, $Z = \frac{\mu + 1.96\sigma - \mu}{\sigma} = 1.96$

When $X = \mu - 1.96\sigma$, $Z = \frac{\mu - 1.96\sigma - \mu}{\sigma} = -1.96$

The transformed distribution would look like below.

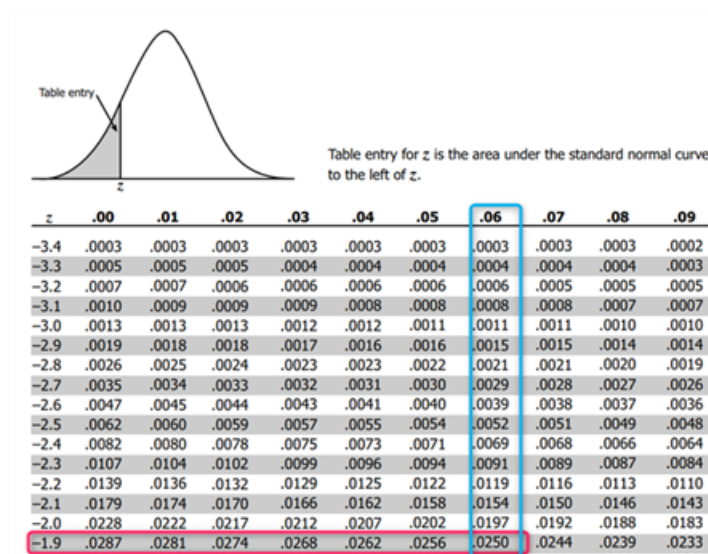


If we look at the Z table for $Z = 1.96$, we will find value as 0.975

Table entry for z is the area under the standard normal curve to the left of z .

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767

If we look at the Z table for $Z = -1.96$, we will find value as 0.025



The area between 0.975 and 0.025 is $0.975 - 0.025 = 0.95$ or 95%. Thus, the value 1.96 was born. It depends on the area we are interested. Here, we were interested in 95% area, so we get $Z = \pm 1.96$

Note The Z table might be left tailed as we just saw or also sometimes right tailed due to symmetrical nature of the curve. This realization is important because when we generalize CI, we will often take right tailed. I used the conventional left tailed table above just to state this explicitly as undoubting readers may miss this point.

CI construction using Pivotal Quantity

From equation 1.14, adding $-\mu$ on both sides of inequalities, we get,

$$\begin{aligned} Pr(-\mu + \mu - 1.96\sigma \leq x - \mu \leq -\mu + \mu + 1.96\sigma) &= 0.95 \\ Pr(-1.96\sigma \leq x - \mu \leq 1.96\sigma) &= 0.95 \end{aligned}$$

And then adding $-x$

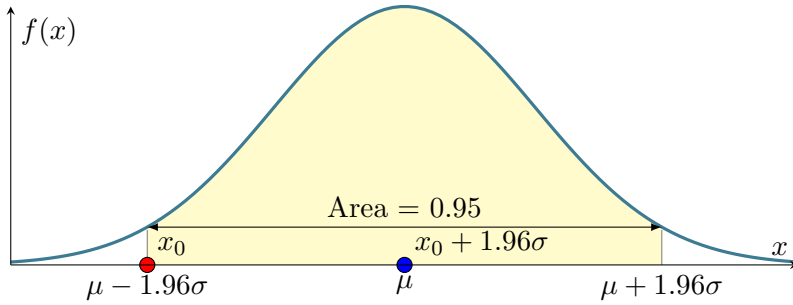
$$\begin{aligned} Pr(-x - 1.96\sigma \leq -x + x - \mu \leq -x + 1.96\sigma) &= 0.95 \\ Pr(-x - 1.96\sigma \leq -\mu \leq -x + 1.96\sigma) &= 0.95 \end{aligned}$$

Multiplying by -1

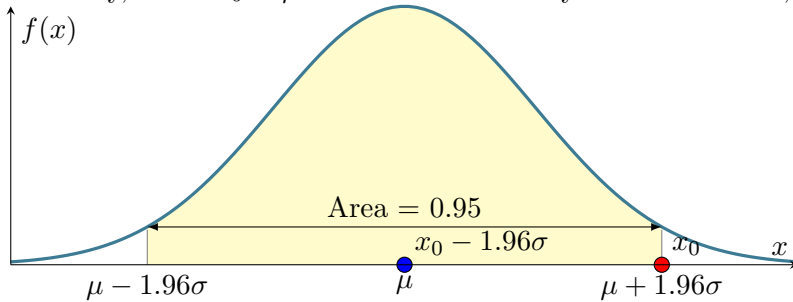
$$\begin{aligned} Pr(x + 1.96\sigma \geq \mu \geq x - 1.96\sigma) &= 0.95 \\ Pr(x - 1.96\sigma \leq \mu \leq x + 1.96\sigma) &= 0.95 \end{aligned} \tag{1.15}$$

Intuitive Proof

Suppose x takes left extreme value within bounds $\mu \pm 1.96\sigma$. That is, $x_0 = \mu - 1.96\sigma$ Then, $\mu = x_0 + 1.96\sigma$



Similarly, when $x_0 = \mu + 1.96\sigma$ then directly we could derive, $\mu = x_0 - 1.96\sigma$



So, as x_0 varies from $\mu - 1.96\sigma$ to $\mu + 1.96\sigma$, implicitly, μ varies from $x_0 + 1.96\sigma$ to $x_0 - 1.96\sigma$. Thus,

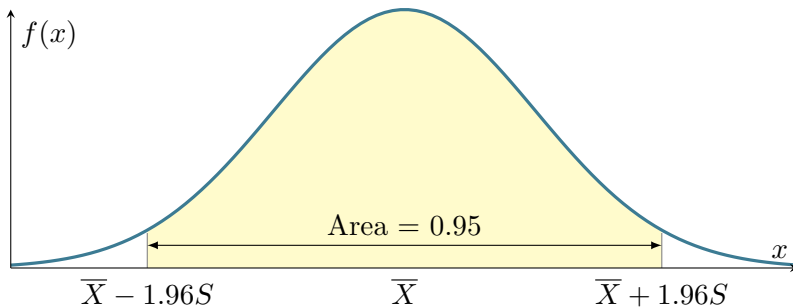
$$Pr(\mu - 1.96\sigma \leq x \leq \mu + 1.96\sigma) = Pr(x - 1.96\sigma \leq \mu \leq x + 1.96\sigma) = 0.95 \quad (1.16)$$

1.2 CI for Sampling Distribution

1.2.1 95% CI as a Corollary

We already have seen, any sampling distribution for sample proportions or sample means, will approach normal distribution, with $\bar{X} \rightarrow \mu$ and $S \rightarrow \frac{\sigma}{\sqrt{n}}$, where μ, σ, n are population mean, population standard deviation, and sample size respectively, when respective conditions¹ are met as per Central Limit theorem (CLT). Note each x is a sample mean.

We thus have a normal distribution like below representing sampling distribution.



¹ $np \geq 10$ and $nq \geq 10$ for sample proportions, $n \geq 30$ for sample means

Then, using equation 1.16, we have,

$$Pr(\bar{X} - 1.96S \leq x \leq \bar{X} + 1.96S) = Pr(x - 1.96S \leq \bar{X} \leq x + 1.96S) = 0.95$$

$$Pr(\mu - 1.96 \frac{\sigma}{\sqrt{n}} \leq x \leq \mu + 1.96 \frac{\sigma}{\sqrt{n}}) = Pr(x - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95 \quad (1.17)$$

Thus, 95% CI for a Sampling distribution would be $(x \pm 1.96 \frac{\sigma}{\sqrt{n}})$

1.2.2 Generalized CI

As hinted in 1.1.4, we will use a right tailed Z table for generalization. We already saw, at Z = -1.96, the area spanned would be 0.025. This could be written as

$$z_{0.025} = -1.96$$

Substituting in 1.17, we get,

$$Pr(x - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95$$

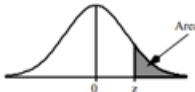
$$Pr(x + z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq x - z_{0.025} \frac{\sigma}{\sqrt{n}}) = 0.95$$

This is kind of counter intuitive. Additive term comes on the LHS. Though one would later discover, $z_{0.025}$ is negative, it could be better if this is not raising any confusion in first place. This is why we use right tailed Z table.

In case of right tailed Z table as below, note, at Z = 1.96, the area spanned is 0.025. Thus we could write it as

$$z_{0.025} = 1.96$$

Normal Curve Areas
Standard normal probability in right-hand tail



z	Second decimal place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

Substituting in 1.17, we get,

$$\begin{aligned} Pr(x - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + 1.96 \frac{\sigma}{\sqrt{n}}) &= 0.95 \\ Pr(x - z_{0.025} \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + z_{0.025} \frac{\sigma}{\sqrt{n}}) &= 0.95 \end{aligned}$$

This is good.

Let α be the desired significance level (which we will learn in hypothesis testing). In our case, it is 5% or $\alpha = 0.05$. Thus, $1 - \alpha = 0.95$ and $\frac{\alpha}{2} = 0.025$

We could then rewrite above equation as,

$$Pr(x - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha \quad (1.18)$$

This is the generalized CI equation where $1 - \alpha$ is called the **confidence coefficient** and $z_{\frac{\alpha}{2}}$ is called the **critical value**

Note Confidence Interval CI indicates not an interval, where population mean is contained 95% of time, but, if one continues to take many such samples and CI for each sample, then 95% of those CIs would contain population mean. We do not know what those CIs are unless we know the population mean and take many such sample sets and their CIs. Once we have taken enough such sample sets (each sample set of size n) calculating CI each time, we could expect that 95% of those CIs have population mean.

1.2.3 When σ is known

In 1.18, we have population standard deviation σ in both end points of the inequalities. Often population parameters are not known in reality. So we have two cases: One when you are lucky enough to know σ and another, you do not know. When you do know, still there are some more parts in play. For example, the more the samples are taken from population, the closer the resulting sampling distribution is to Normal (or Normal approximation is becoming better), so when do you say, sample size n is good enough? This depends on various conditions.

1. If we sample from population whose distribution is itself normal, then even small sample size $n \geq 5$ would suffice because our sampling distribution easily approximates to Normal. Our current CI equation holds good.

$$Pr(x - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

2. If we sample from population whose distribution is not normal but symmetric, unimodal and of the continuous type, then as per Central limit theorem (CLT), sample size $n \geq 30$ should be adequate generally as this would result in sampling distribution becoming almost normal so our equation could still be approximately good. That is,

$$Pr(x - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) \approx 1 - \alpha \quad (1.19)$$

3. If distribution is non normal and also highly skewed, even above approximation would not work. In that case, it would be safer to use certain nonparametric methods for finding a CI for the median of the distribution.

1.2.4 When σ is not known

This is often the case in reality. In this case, depending on certain conditions like above, we could use student's t distribution². The t distribution looks like normal, except the tails are bigger, and also depends on degrees of freedom (which usually is $n - 1$). The proof is exhaustive, so we will take at face value for now (and prove in future if time permits)

1. If we sample from population whose distribution is itself normal, and if sample size $n \leq 30$, then our CI equation would be,

$$Pr(x - t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq x + t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}) = 1 - \alpha \quad (1.20)$$

where $t_{\frac{\alpha}{2},(n-1)}$ is the t value for probability area $\frac{\alpha}{2}$, for degrees of freedom $(n - 1)$ from corresponding right tailed t table.

2. If we sample from population whose distribution is itself normal, and if sample size $n > 30$, then our t distribution would already be almost equal to normal (and resulting sampling distribution would be normal) so we could use as below,

$$Pr(x - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq x + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}) = 1 - \alpha \quad (1.21)$$

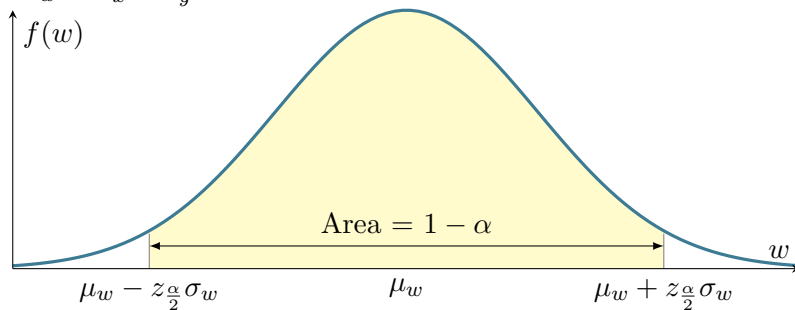
3. If we sample from population whose distribution is not normal but symmetric, unimodal and of the continuous type, and sample size $n \leq 30$, we get approximate CI as below.

$$Pr(x - t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq x + t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}) \approx 1 - \alpha \quad (1.22)$$

4. If distribution is non normal and also highly skewed, even above approximation would not work. In that case, it would be safer to use certain nonparametric methods for finding a CI for the median of the distribution.

1.2.5 CI for difference between two means

This section is heavily inspired by Robert et al. [2], and I have tried to articulate in my style to my understanding. Suppose that we are interested in comparing two approximately normal sampling distributions described by random variables $\bar{X} = N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$ and $\bar{Y} = N(\mu_{\bar{y}}, \sigma_{\bar{y}}^2)$, created from population distributions described by random variables $X(\mu_x, \sigma_x^2)$ and $Y(\mu_y, \sigma_y^2)$. Note that \bar{X} represents collection of sample means from sampled sets sampled from X and similarly for \bar{Y} . Since both \bar{X} and \bar{Y} are normally distributed, and assuming both are independent to each other, the distribution $W = \bar{X} - \bar{Y}$ would be again a normal distribution $W(\mu_w, \sigma_w^2)$, where $\mu_w = \mu_{\bar{x}} - \mu_{\bar{y}}$ and $\sigma_w^2 = \sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2$ as proved in 3.0.1



²<http://pages.wustl.edu/montgomery/articles/2757>

Since W is a normal distribution now, we have the confidence interval as follows directly following equation 1.14

$$\begin{aligned}
Pr(\mu - z_{\frac{\alpha}{2}}\sigma \leq x \leq \mu + z_{\frac{\alpha}{2}}\sigma) &= 1 - \alpha \\
Pr(\mu_w - z_{\frac{\alpha}{2}}\sigma_w \leq W \leq \mu_w + z_{\frac{\alpha}{2}}\sigma_w) &= 1 - \alpha \\
Pr(-z_{\frac{\alpha}{2}}\sigma_w \leq W - \mu_w \leq z_{\frac{\alpha}{2}}\sigma_w) &= 1 - \alpha \\
Pr(-z_{\frac{\alpha}{2}} \leq \frac{W - \mu_w}{\sigma_w} \leq z_{\frac{\alpha}{2}}) &= 1 - \alpha \\
Pr(-z_{\frac{\alpha}{2}} \leq \frac{W - \mu_w}{\sigma_w} \leq z_{\frac{\alpha}{2}}) &= 1 - \alpha \\
Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_{\bar{x}} - \mu_{\bar{y}})}{\sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2}} \leq z_{\frac{\alpha}{2}}\right) &= 1 - \alpha \\
Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_{\bar{x}} - \mu_{\bar{y}})}{\sqrt{\frac{\sigma_{\bar{x}}^2}{n} + \frac{\sigma_{\bar{y}}^2}{m}}} \leq z_{\frac{\alpha}{2}}\right) &= 1 - \alpha \tag{1.23}
\end{aligned}$$

where $Z = \frac{W - \mu_w}{\sigma_w}$ would be the "standardized" normal distribution $N(0,1)$, n and m are sample set sizes of $X(\mu_x, \sigma_x)$ and $Y(\mu_y, \sigma_y)$ respectively.

Assuming σ unknown

Most of the times in reality, the population paramters are not known. So when the sample sizes n, m are sufficiently large, we could use sample SDs ($s_{\bar{x}}, s_{\bar{y}}$) in place of (σ_x, σ_y) .

$$Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{(\bar{X} - \bar{Y}) - (\mu_{\bar{x}} - \mu_{\bar{y}})}{\sqrt{\frac{s_{\bar{x}}^2}{n} + \frac{s_{\bar{y}}^2}{m}}} \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

And also rewriting, to find CI for $(\mu_{\bar{x}} - \mu_{\bar{y}})$, we get,

$$Pr\left((\bar{X} - \bar{Y}) - z_{\frac{\alpha}{2}}s_w \leq (\mu_{\bar{x}} - \mu_{\bar{y}}) \leq (\bar{X} - \bar{Y}) + z_{\frac{\alpha}{2}}s_w\right) \approx 1 - \alpha \tag{1.24}$$

where, $s_w = \sqrt{\frac{s_{\bar{x}}^2}{n} + \frac{s_{\bar{y}}^2}{m}}$, and n, m are large.

When n, m are small

We would then use student's t distribution as suggested by **Welch and Aspin**. The proof is currently beyond the scope so we take it at face value.

$$Pr\left((\bar{X} - \bar{Y}) - t_{(\frac{\alpha}{2}, r)}s_w \leq (\mu_{\bar{x}} - \mu_{\bar{y}}) \leq (\bar{X} - \bar{Y}) + t_{(\frac{\alpha}{2}, r)}s_w\right) \approx 1 - \alpha \tag{1.25}$$

where r is degrees of freedom. Since two distributions are involved, calculating r is complicated. It is given as follows:

$$r = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2} \tag{1.26}$$

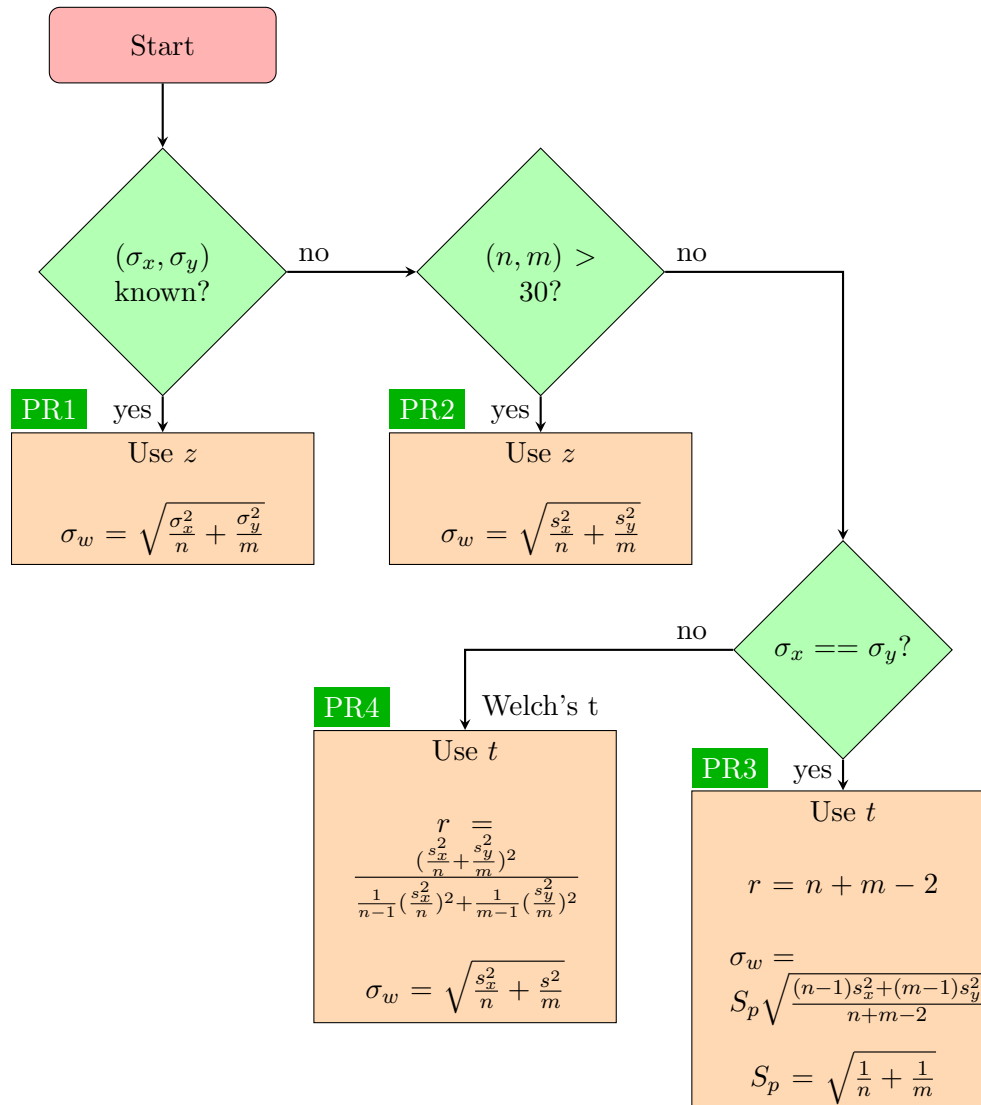
Protection when $\sigma_x = \sigma_y$

Since we do not know σ_x, σ_y , it might be that they are also equal. If they happen to be equal, r could be proven as below.

$$r = (n - 1) + (m - 1) = n + m - 2$$

The equation 1.26 protects in the sense that, the r value from that is lesser than above equation, so t value is higher, or t distribution of wider variance assumed, thus being conservative. Some texts simply also take $r = \min(n - 1, m - 1)$ as conservative approach.

A visual summary



1.2.6 CI for difference between two proportions

Suppose that we are interested in comparing two approximately normal sampling distributions described by random variables $\frac{Y_1}{n_1} = N\left(p_1, \frac{p_1 q_1}{n_1}\right)$ and $\frac{Y_2}{n_2} = N\left(p_2, \frac{p_2 q_2}{n_2}\right)$, created from population distributions which are Bernoulli distributions.

Note that Y_1 represents the sum of *successes* in a sample set, and thus $\frac{Y_1}{n_1}$ represents sample proportions. For example, for any k th sample set of $\frac{Y_1}{n_1}$, we calculate sample proportion statistic,

$\frac{Y_{1k}}{n_1} = \frac{1}{n} \sum_{i=1}^n Y_{1ki}$, where Y_{1ki} is i th sample in k th sample set of sampling distribution described by $\frac{Y_1}{n_1}$. Similarly for $\frac{Y_2}{n_2}$.

We could then rewrite 1.23 as below

$$Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{\left(\frac{Y_1}{n_1} - \frac{Y_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

In case you are wondering about the parameters inside, say $W = \frac{Y_1}{n_1} - \frac{Y_2}{n_2}$, then

$$\mu_w = \mu_{y_1/n_1} - \mu_{y_2/n_2} = p_1 - p_2$$

$$\sigma_w^2 = \sigma_{y_1/n_1}^2 + \sigma_{y_2/n_2}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2} \therefore \sigma_w = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Assuming σ unknown

Most of the times in reality, the population parameters are not known. So when the sample sizes n, m are sufficiently large, we could use sample statistics $\left(\frac{\hat{p}_1 \hat{q}_1}{n_1}, \frac{\hat{p}_2 \hat{q}_2}{n_2}\right)$ in place of $\left(\frac{p_1 q_1}{n_1}, \frac{p_2 q_2}{n_2}\right)$. This results in further approximation of our confidence intervals. Thus when a sample is observed, we have statistics

$$\hat{p}_1 = \frac{y_1}{n_1}, \hat{q}_1 = 1 - \frac{y_1}{n_1}, \hat{p}_2 = \frac{y_2}{n_2}, \hat{q}_2 = 1 - \frac{y_2}{n_2},$$

Thus we could rewrite further as,

$$Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}} \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha \quad (1.27)$$

When n, m are small

Currently I do not have an answer for this question and could not find online. Raised a ticket(?) [here](#)

Chapter 2

Examples

2.1 Deep Examples

2.1.1 Confidence Intervals for Sampling Proportions

Create Population

Let us create a population of 10000 balls, with 60% yellow balls. Programmatically, our population contains 1s and 0s, 1 indicating yellow.

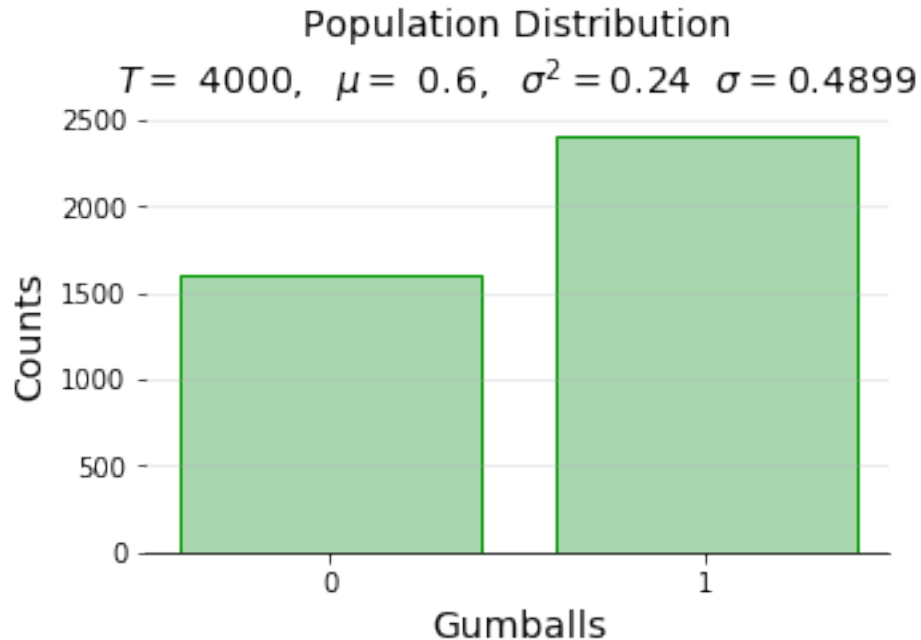
```
In[1]: %matplotlib inline
import matplotlib.pyplot as plt
from SDSPSM import get_metrics, drawBarGraph
from ci_helpers import create_bernoulli_population

T = 4000 # total size of population
p = 0.6 # 60% has yellow balls

# create population
population, population_freq = create_bernoulli_population(T,p)

# population metrics
mu, var, sigma = get_metrics(population)

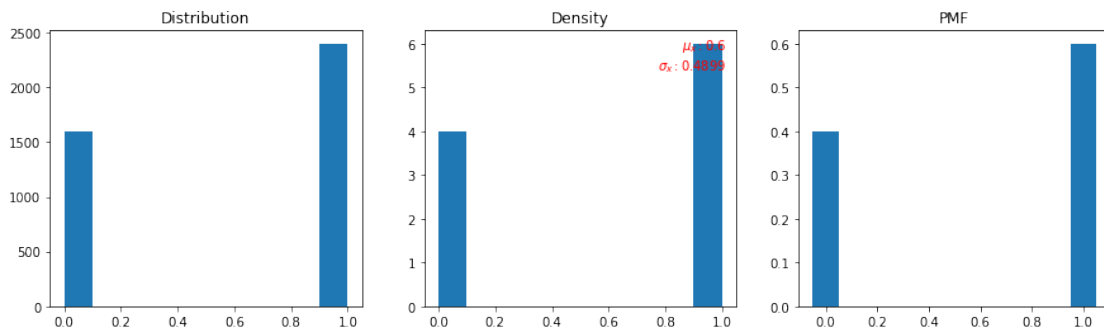
# visualize
fig, (ax1) = plt.subplots(1,1, figsize=(5,3))
drawBarGraph(population_freq, ax1, [T, mu, var, sigma], 'Population
Distribution', 'Gumballs', 'Counts', xmin=0)
plt.show()
```



Deriving and visualizing the probability Mass function (the intermediate density function, where total area of bars will be 1, is just for fitting normal continuous approximation later)

```
In[2]: from ci_helpers import mini_plot_SDSP

fig, (ax1,ax2,ax3) = plt.subplots(1,3,figsize=(15,4))
mini_plot_SDSP(population, ax1,ax2,ax3, norm_off=True)
plt.show()
```



Sampling from the Population

Let us sample from population, N no of times, each time with sample set of size n . If $np \geq 30$ and $nq \geq 30$, the resulting sampling distribution should be approximately normal. Remember, for Population described by random variable Y , we describe the sampling distribution by

for any sample set k , sample mean is $\widehat{Y}_k = \frac{1}{n} \sum_{i=1}^n Y_{ki}$

$$\text{Random Variable } \widehat{p} = \widehat{Y} = \widehat{Y}_1, \widehat{Y}_2, \dots, \widehat{Y}_k \dots \widehat{Y}_N \quad (2.1)$$

$$\mu_{\widehat{p}} = \mu(\widehat{Y})$$

$$\sigma_{\widehat{p}} = \sigma(\widehat{Y})$$

where the hat $\widehat{}$ indicates the statistical outcome. And statistically by CLT,

$$\mu_{\widehat{p}} \approx 0.6 = \mu = p$$

$$\sigma_{\widehat{p}} \approx 0.0693 \approx \frac{0.4898}{\sqrt{50}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}} \quad (2.2)$$

Note we have sampled WITH REPLACEMENT, so the samples are independent. If you sample without replacement, you need to factor in FPC (finite population correction) for each sample set's SD.

```
In[3]: from ci_helpers import sample_with_CI
        from random import seed

        N = 100
        n = 50

        #seed(0)

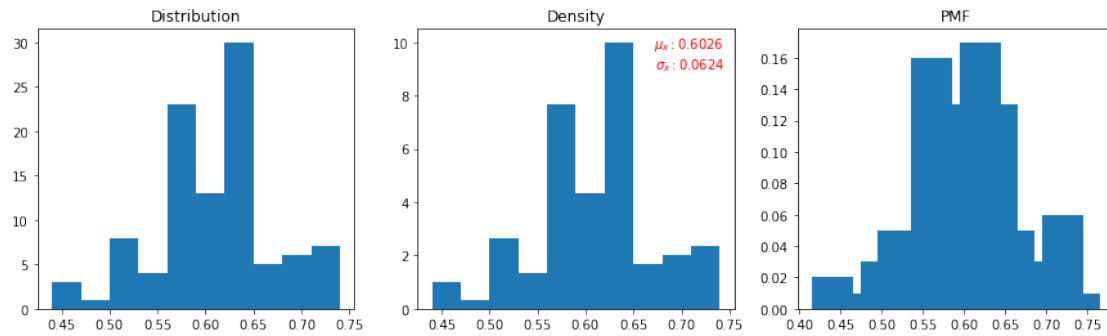
        # sample from population
        Y_mean_list, CI_list = sample_with_CI(N, n, population, sigma=sigma, mode='z')

        # sample metrics
        mu, var, sigma = get_metrics(Y_mean_list)

        # visualize
        fig, (ax1,ax2,ax3) = plt.subplots(1,3,figsize=(15,4))
        mini_plot_SDSP(Y_mean_list,ax1,ax2,ax3,width=0.05, norm_off=True)

        from IPython.display import display, Math
        display(Math(r'\mu_{\{\hat{p}\}}:{} \ \ \ \ \sigma_{\{\hat{p}\}}:{}'.format(mu, sigma)))
```

$$\mu_{\widehat{p}} : 0.6026 \quad \sigma_{\widehat{p}} : 0.0624$$



When σ is known

For each of above sample set of size 'n', let us calculate confidence interval using population SD σ as below. 1.96 is from Z transformation for 95% confidence interval, like we saw earlier in our theoretical section.

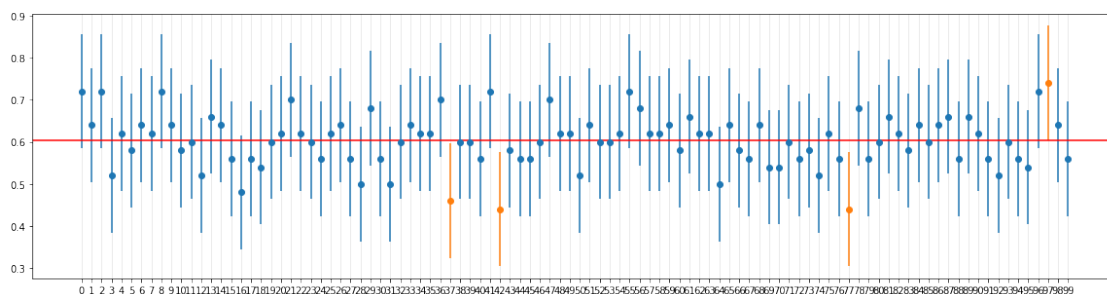
$$CI = Y \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad (2.3)$$

```
In[4]: from ci_helpers import plot_ci_accuracy_1

fig, ax = plt.subplots(1,1, figsize=(20,5))

plot_ci_accuracy_1(ax, CI_list, mu)
plt.show()
```

CI containing pop.mean:96.0%



As expected we observe that out of all CIs above, 95% of them or above contain population mean.

When σ is not known

For each sample mean \bar{X}_k calculated, the confidence interval is calculated as below. Note, the constant value t_{n-1} depends on degrees of freedom (n-1).

$$CI = Y \pm t_{n-1} \frac{S_k}{\sqrt{n}} \quad (2.4)$$

Hope you noted. This time, for each sample mean, we also calculate unbiased sample variable of that set (that is, divided by $n-1$), and use that for calculating M_k . We sample again, because, for each sample, this time, we calculate CI using t distribution.

t value for 95% CI:

Degrees of Freedom $df = n - 1$. For 95% confidence level, the confidence coefficient, $1 - \alpha = 1 - 0.05 = 0.95$.

To calculate t in python, we simply need to pass, $(1 - \alpha, df)$. A sample calculation shown below for sample size $n = 10$

```
In[5]: from scipy import stats
print(stats.t.ppf(1-0.025, 10-1))
```

2.2621571627409915

Now to our sampling distribution. Note, we are getting an approximate normal distribution.

```
In[6]: from ci_helpers import sample_with_CI

N = 100
n = 50

#seed(0)

# sample from population, this time in t mode,
# so CI intervals are calculated with t value 2.093
Y_mean_list, CI_list = sample_with_CI(N, n, population, sigma=sigma, mode='t')

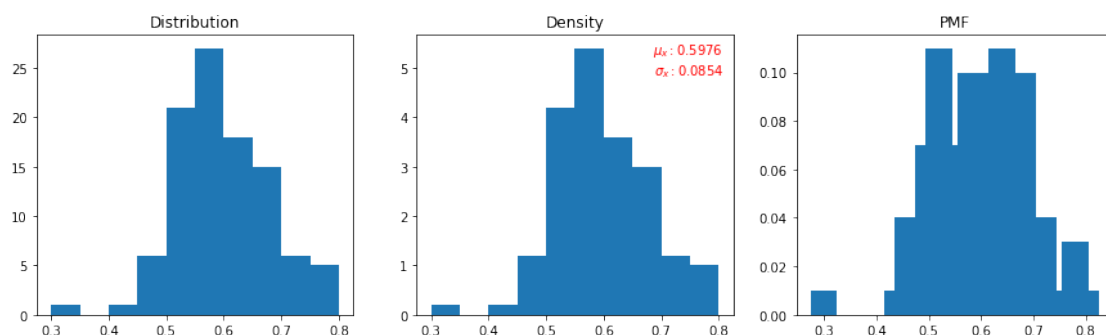
# sample metrics
mu, var, sigma = get_metrics(Y_mean_list)

# visualize
fig, (ax1,ax2,ax3) = plt.subplots(1,3,figsize=(15,4))
mini_plot_SDSP(Y_mean_list,ax1,ax2,ax3,width=0.05, norm_off=True)

from IPython.display import display, Math
display(Math(r'\mu_{\hat{p}}: {} \ \ \ \ \sigma_{\hat{p}}: {}'.format(mu, sigma)))

plt.show()
```

$$\mu_{\hat{p}} : 0.5976 \quad \sigma_{\hat{p}} : 0.0854$$

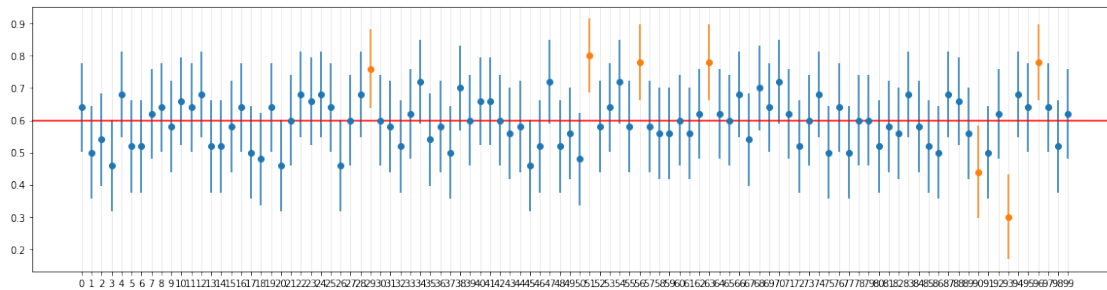


```
In[7]: from ci_helpers import plot_ci_accuracy_1

fig, ax = plt.subplots(1,1, figsize=(20,5))

plot_ci_accuracy_1(ax, CI_list, mu)
plt.show()
```

CI containing pop.mean:93.0%



Generally we should get more than 95% as above. Above result just means, if we take a sampling size, and calculate CI, and do that 100 times, about 95 times our CI would contain population mean, and our result gave 97 times. We could expect at least 95% most of the time. But can we get any idea, how that "success" of getting population mean in our CI, 95% of time, depends on sample size? We get it, greater the sample size, better, but how it would be? Let us take our simulation to next scale as below, trying with various experiment and sample sizes.

Digging deeper 1

What if I use Z distribution and unbiased sample SD even for CI? What happens when I use t distribution but population SD for CI? We will find out what happens in such cases below.

Environment:

1. Population size T, fixed
2. Sample size n, varied
3. Experiment size N, varied
4. Sampling with or without replacement, varied.

Applied methods:

1. Z distribution and population SD
2. Z distribution and unbiased sample SD
3. T distribution and population SD
4. T distribution and unbiased sample SD

Note, in case of sampling without replacement, each sample SD is corrected with FPC (Finite Population Correction)

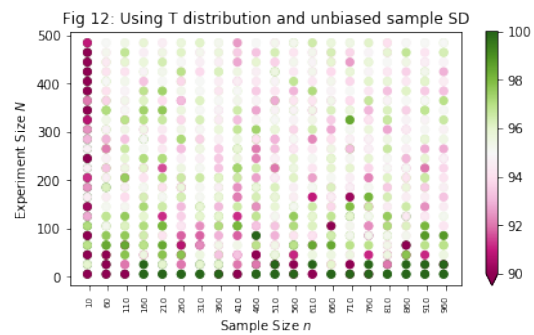
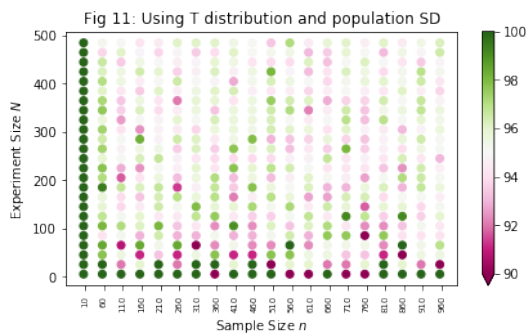
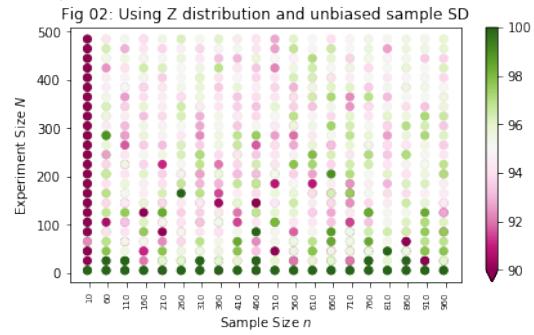
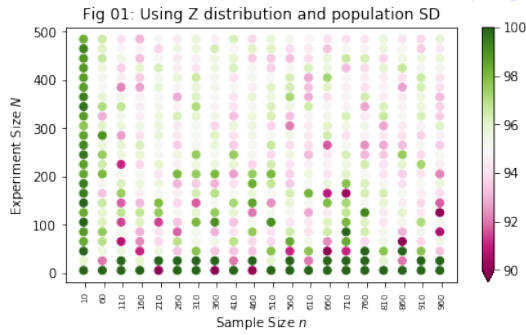
```
In[8]: from ci_helpers import plot_summary

max_sample_size = int(T/4) # 25% of total population
N_list = range(5,500,20)
```

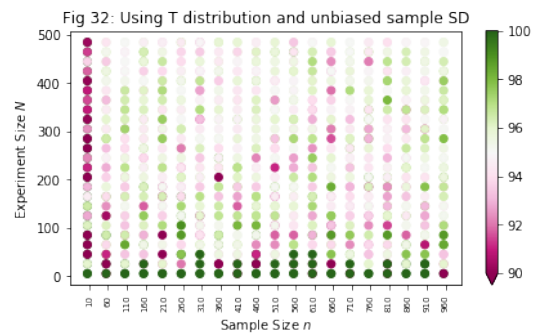
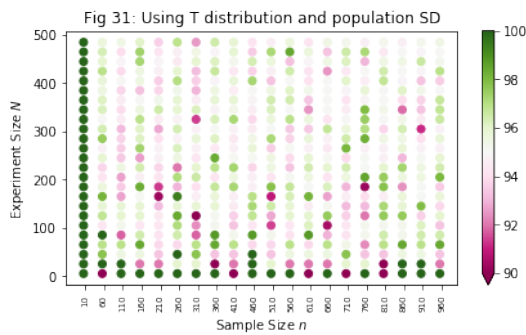
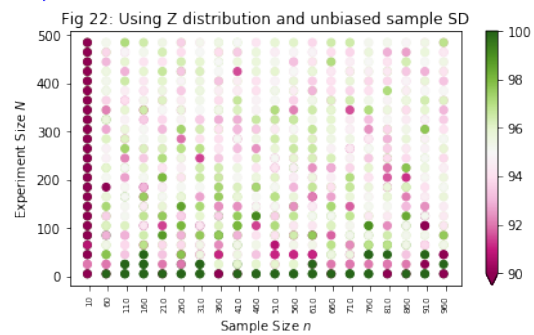
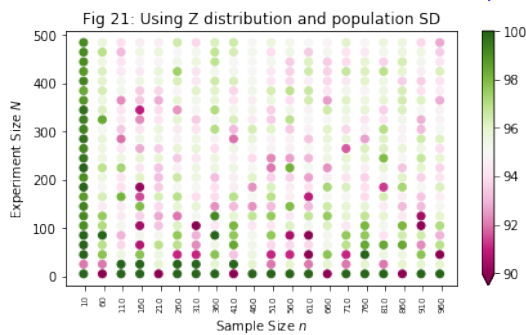
```
n_list = range(5,max_sample_size,50) # different sample sizes

plot_summary(population, N_list, n_list)
```

Sampling Without Replacement



Sampling With Replacement



Note that, as per color gradient used, lighter the dots, nearer they are to 95%. And if green they are above 95%. And if pink, they are below 95%. So more the green dots or lighter dots, the better, the CI performance.

1. Compared to graphs using sample SDs on right hand side, the graphs using population SDs on left hand side, has more dots that are green and lighter indicating better CI performance on LHS. This is especially very pronounced, when sample sizes are small (observe dark dots at $n = 10$). LHS almost always have green dots at $n = 10$ while RHS has mostly pinky dots.
2. For a common SD usage, there is not much a difference between using Z or t distribution when $n \geq 30$. For eg, compare figures 01 and 11 both using population SD. Or compare 02 and 12 both using sample SD.
3. Comparing figures 01 and 11 at $n = 10$ we observe, figure 11 performs better (more darker green dots). So when you know σ , and if $n < 30$ using Z distribution is better.
4. Comparing figures 02 and 12 at $n = 10$ we observe, figure 12 performs better (lighter pink dots). So when you do not know σ and if $n < 30$, using T distribution with unbiased sample SD is better.
5. Similar observation also applies to sampling with replacement.

Though the limit 30 is not obvious from above graphs, this number has been arrived at by statisticians after extensive research

Warning

The CI for proportions have been always blotchy. Though above formula are straight forward, they have been proven ineffective, effectively by Brown et al. [1]. When you use CI for proportions problem in a practical scenario do use the alternatives provided there. In a nutshell, for smaller sizes, $n < 40$, Wilson or the equal-tailed Jeffreys prior interval are recommended. For larger n, the Wilson, the Jeffreys and the Agresti–Coull intervals are all comparable, and the Agresti–Coull interval is the simplest to present.

2.1.2 Confidence Intervals for Sample Means

Create Population

Let Y be the random variable indicating temperature over a distribution of certain values. If limiting values are say, 0 deg C to 40 deg C, our population would thus look like this: [23, 13, 35, 50, 10, 2, 5, 0, 33, \dots , 21] Unlike Sample proportions, we do not know or designate any proportion of temperatures in this example, but we know the mean and variance by simply calculating all values in the distribution. These would be our population parameters.

Population mean $\mu = \mu_y$
Population variance $\sigma^2 = \sigma_y^2$

```
In[9]: %matplotlib inline
        from math import floor
        import matplotlib.pyplot as plt
        from random import random, seed, shuffle
        from SDSPSM import get_metrics, drawBarGraph, getPopulationStatistics
        from ci_helpers import createRandomPopulation

        seed(0)

        popMin = 1 # Min population
```

```

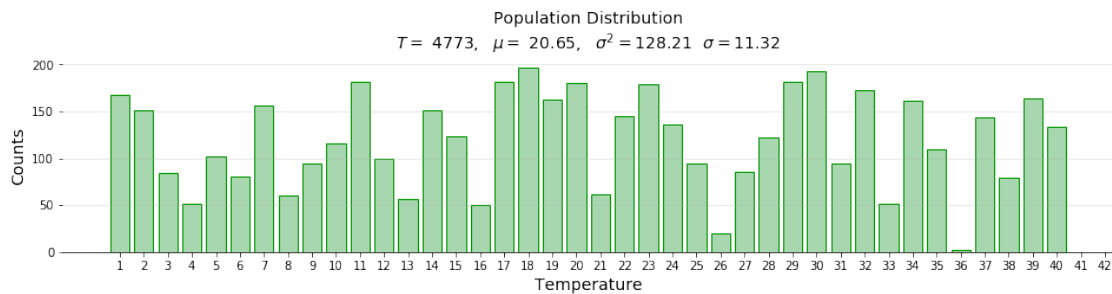
popMax = 40 # Max population
freqMax = 200 # freq of any set of population (for eg, no of occurrences of temperatures
at 25 deg C)

population, population_freq = createRandomPopulation(popMax - popMin + 1, freqMax)

N, mu, var, sigma = getPopulationStatistics(population_freq, popMin)

#visualize
fig, (ax1) = plt.subplots(1,1, figsize=(16,3))
drawBarGraph(population_freq, ax1, [N, mu, var, sigma], 'Population
Distribution', 'Temperature', 'Counts')
plt.show()

```



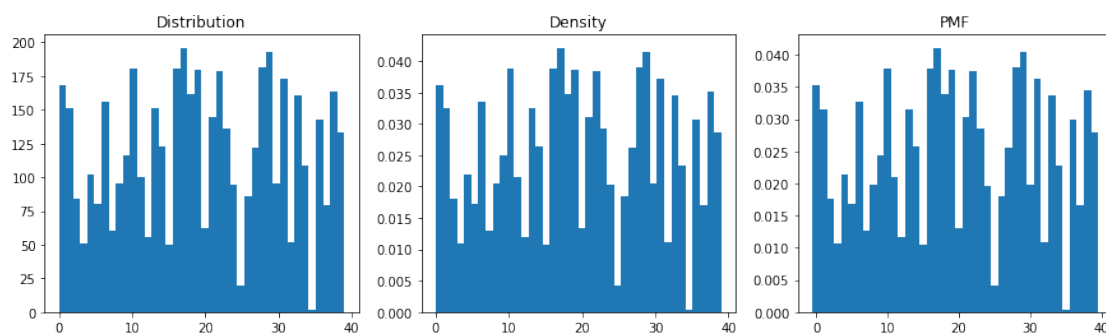
Let us visualize the density function and PMF as usual.

```

In[10]: from ci_helpers import mini_plot_SDSM

fig, (ax1,ax2,ax3) = plt.subplots(1,3,figsize=(15,4))
mini_plot_SDSM(population, ax1, ax2, ax3, popMax, width=1)
plt.show()

```



Sampling from the Population

Let us sample from above population, N no of times, each time with sample set of size n . If $n > 30$, the resulting sampling distribution should be approximately normal (always if population itself was normally distributed)

Remember, for Population described by random variable Y , we describe the sampling distribution of sample means by

$$\mu_{\bar{Y}} = \mu(\widehat{Y}) \quad (2.5)$$

$$\sigma_{\bar{Y}} = \sigma(\widehat{Y})$$

where the $\widehat{\cdot}$ indicates the statistical outcome. And statistically by CLT,

$$\mu_{\bar{Y}} = 19.4 \approx 20 = \mu$$

$$\sigma_{\bar{Y}} \approx 1.52 \approx \frac{11.32}{\sqrt{50}} = \frac{\sigma}{\sqrt{n}} \quad (2.6)$$

\bar{Y} is called the sample means which is a random variable.

```
In[11]: from ci_helpers import sample_with_CI
        from random import seed

        N = 100
        n = 50

        #seed(0)

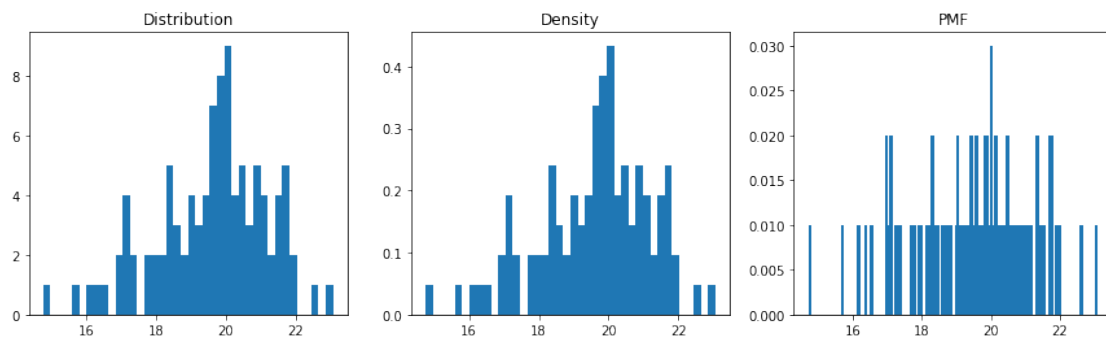
        # sample from population
        Y_mean_list, CI_list = sample_with_CI(N, n, population, sigma=sigma, mode='z')

        # sample metrics
        mu, var, sigma = get_metrics(Y_mean_list)

        # visualize
        fig, (ax1,ax2,ax3) = plt.subplots(1,3,figsize=(15,4))
        mini_plot_SDSM(Y_mean_list, ax1, ax2, ax3, popMax, width=0.1)

        from IPython.display import display, Math
        display(Math(r'\mu_{\hat{p}}:{} \ \ \ \ \sigma_{\hat{p}}:{}'.format(mu, sigma)))
```

$$\mu_{\hat{p}} : 19.5912 \quad \sigma_{\hat{p}} : 1.5865$$



Ok I get it, the resulting distribution and density functions look abnormal (ugly, slightly normal). Try increasing experiment size N , and you will see much better approximation of normal distribution. We had to stick with $N=100$ because we have to see how CI from each sample mean performs, so bear with me here.

When σ is known

For each of above sample set of size 'n', let us calculate confidence interval using population SD σ as below. 1.96 is from Z transformation for 95% confidence interval, like we saw earlier in our theoretical section.

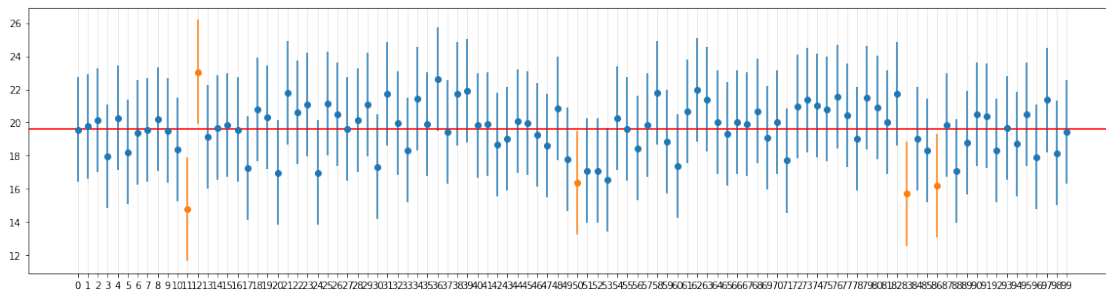
$$CI = Y \pm 1.96 \frac{\sigma}{\sqrt{n}} \quad (2.7)$$

```
In[12]: from ci_helpers import plot_ci_accuracy_1

        fig, ax = plt.subplots(1,1, figsize=(20,5))

        plot_ci_accuracy_1(ax, CI_list, mu)
        plt.show()
```

CI containing pop.mean:95.0%



When σ is not known

When we do not know population SD

Just like earlier, for each sample mean \bar{X}_k calculated, the confidence interval is calculated as below. Note, the constant value t_{n-1} depends on degrees of freedom (n-1).

$$CI = Y \pm t_{n-1} \frac{S_k}{\sqrt{n}} \quad (2.8)$$

```
In[13]: from ci_helpers import sample_with_CI

        N = 100
        n = 50

        #seed(0)

        # sample from population, this time in t mode,
        # so CI intervals are calculated with t value 2.093
        Y_mean_list, CI_list = sample_with_CI(N, n, population, sigma=sigma, mode='t')

        # sample metrics
        mu, var, sigma = get_metrics(Y_mean_list)

        # visualize
        fig, (ax1,ax2,ax3) = plt.subplots(1,3,figsize=(15,4))
```

```

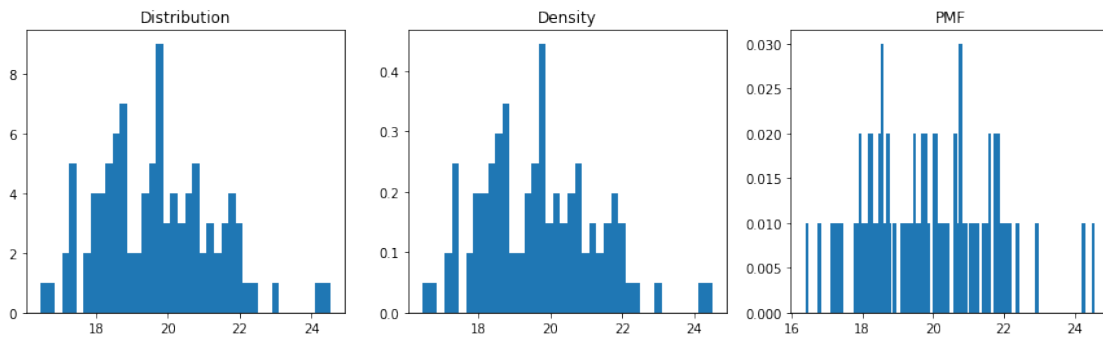
mini_plot_SDSM(Y_mean_list, ax1, ax2, ax3, popMax, width=0.1)

from IPython.display import display, Math
display(Math(r'\mu_{\hat{p}}:{} \ \ \ \ \sigma_{\hat{p}}:{}'.format(mu, sigma)))

plt.show()

```

$$\mu_{\hat{p}} : 19.6824 \quad \sigma_{\hat{p}} : 1.5962$$



```

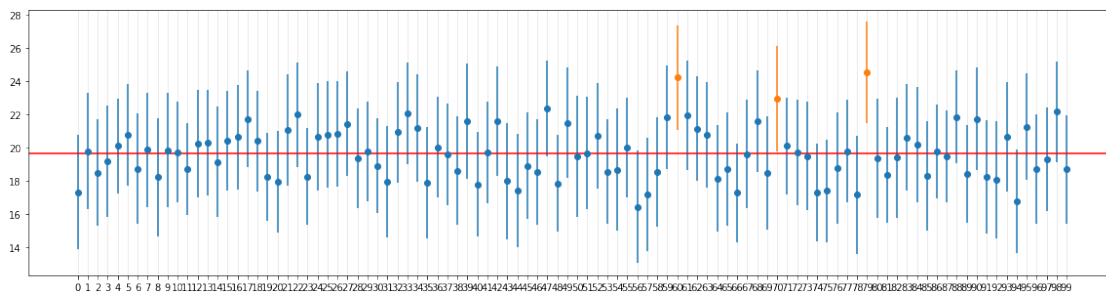
In[14]: from ci_helpers import plot_ci_accuracy_1

fig, ax = plt.subplots(1,1, figsize=(20,5))

plot_ci_accuracy_1(ax, CI_list, mu)
plt.show()

```

CI containing pop.mean:97.0%



Digging deeper 2

What if I use Z distribution and unbiased sample SD even for CI? What happens when I use t distribution but population SD for CI? We will find out what happens in such cases below.

Environment:

1. Population size T, fixed
2. Sample size n, varied

3. Experiment size N , varied
4. Sampling with or without replacement, varied.

Applied methods:

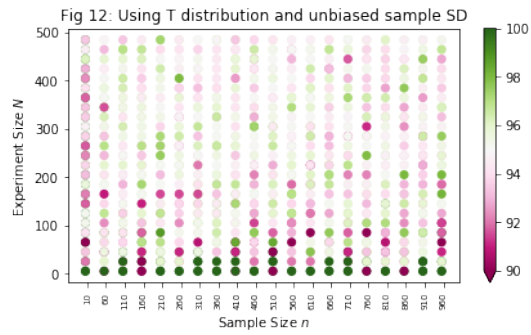
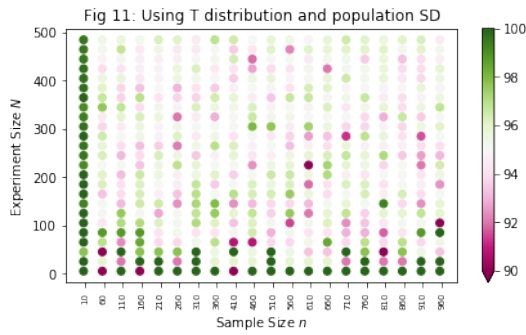
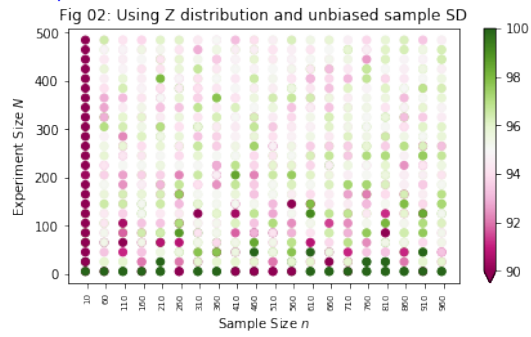
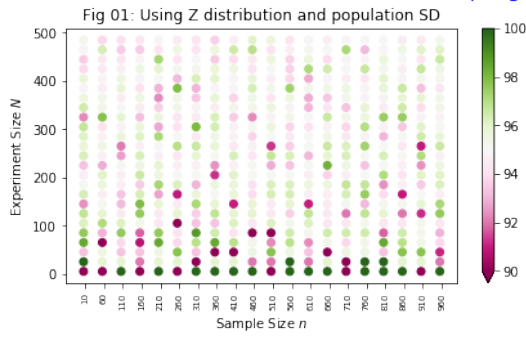
1. Z distribution and population SD
2. Z distribution and unbiased sample SD
3. T distribution and population SD
4. T distribution and unbiased sample SD

Note, in case of sampling without replacement, each sample SD is corrected with FPC (Finite Population Correction)

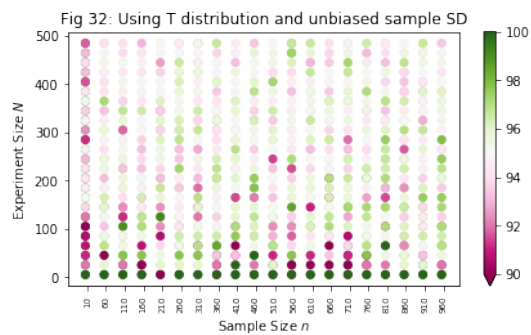
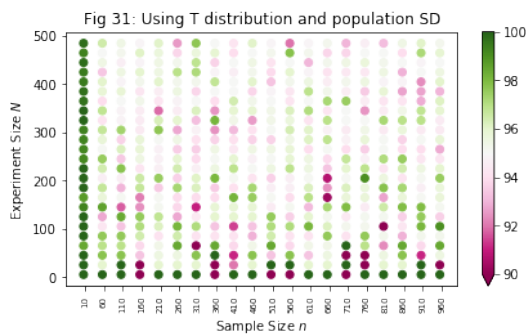
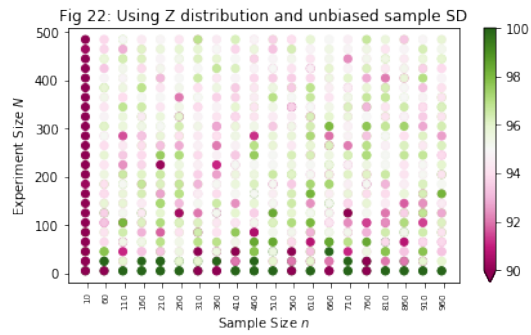
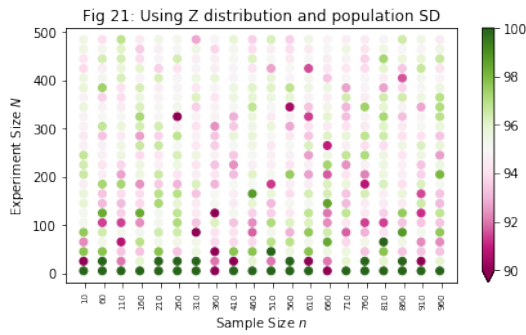
```
In[15]: max_sample_size = int(T/4) # 25% of total population
        N_list = range(5,500,20)
        n_list = range(5,max_sample_size,50) # different sample sizes

        plot_summary(population, N_list, n_list)
```

Sampling Without Replacement



Sampling With Replacement



Note that, as per color gradient used, lighter the dots, nearer they are to 95%. And if green they are above 95%. And if pink, they are below 95%. So more the green dots or lighter dots, the better, the CI performance.

1. Compared to graphs using sample SDs on right hand side, the graphs using population SDs on left hand side, has more dots that are green and lighter indicating better CI performance on LHS. This is especially very pronounced, when sample sizes are small (observe dark dots at $n = 10$). LHS almost always have green dots at $n = 10$ while RHS has mostly pinky dots.
2. For a common SD usage, there is not much a difference between using Z or t distribution when $n \geq 30$. For eg, compare figures 01 and 11 both using population SD. Or compare 02 and 12 both using sample SD.
3. Comparing figures 01 and 11 at $n = 10$ we observe, figure 11 performs better (more darker green dots). So when you know σ , and if $n < 30$ using Z distribution is better.
4. Comparing figures 02 and 12 at $n = 10$ we observe, figure 12 performs better (lighter pink dots). So when you do not know σ and if $n < 30$, using T distribution with unbiased sample SD is better.
5. Similar observation also applies to sampling with replacement.

Though the limit 30 is not obvious from above graphs, this number has been arrived at by statisticians after extensive research.

Yes, the inferences are same as Section 2.1.1 except that the differences are much more clearer in this case. For eg, compare figures 02 and 12 at $n = 10$. It is very clear now, why figure 12 (using t distribution) is far better at lower sample sizes.

2.2 Shallow Examples

2.2.1 σ Known, Population Normal, Low Sample Size

Let X equal the length of life of a 60-watt light bulb marketed by a certain manufacturer. Assume that the distribution of X is $N(\mu, 1296)$. If a random sample of $n = 27$ bulbs is tested until they burn out, yielding a sample mean of $x = 1478$ hours, find 95% confidence interval for μ .

Solution: Here, its given that the population is Normal and also its population SD σ . So we could use equation 1.18 right away. Given

$$\sigma^2 = 1296 \therefore \sigma = 36,$$

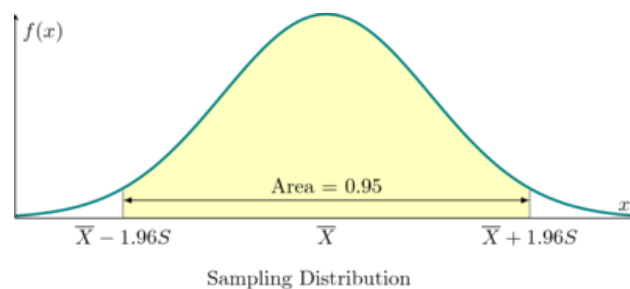
$$x = 1478, 1 - \alpha = 0.95,$$

$$z_{\frac{\alpha}{2}} = z_{0.025} = 1.96, n = 27 \geq 5$$

Though sample size is < 30 , the population distribution is given as normal already. Thus, our sampling distribution would still be a normal distribution as below with 95% confidence interval area.

The tikzmagic extension is already loaded. To reload it, use:

```
%reload.ext tikzmagic
```



We already know, in this sampling distribution, the mean $\bar{X} \rightarrow \mu$ and SD $S \rightarrow \frac{\sigma}{\sqrt{n}}$. Thus as we have already derived earlier,

$$\begin{aligned} Pr(\bar{X} - 1.96S \leq x_0 \leq \bar{X} + 1.96S) &= 1 - \alpha \\ Pr(x_0 - 1.96S \leq \bar{X} \leq x_0 + 1.96S) &= 1 - \alpha \\ Pr\left(x_0 - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq x_0 + 1.96\frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\ \implies Pr\left(1478 - 1.96\frac{36}{\sqrt{27}} \leq \mu \leq 1478 + 1.96\frac{36}{\sqrt{27}}\right) &= 0.95 \\ Pr(1478 - 13.58 \leq \mu \leq 1478 + 13.58) &= 0.95 \\ Pr(1464.42 \leq \mu \leq 1491.58) &= 0.95 \end{aligned}$$

Thus the 95% CI intervals are [1464.42, 1491.58]. This does not mean, μ is inside this interval 95% of the time. But simply, if we are to take many such samples and their CIs, 95% of those CIs would contain μ . We do not know what those CIs would be because we do not know the real μ .

2.2.2 σ Known, Population not Normal, High Sample Size

The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing 120 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a 92% confidence interval for the mean assembly time.

Solution:

Given $n = 120$ which is > 30 . The measurement in population is *mean amount of time* which is *continuous*. Due to CLT, the resulting sampling distribution of sample means from all sample sets of size $n = 120$ would result in a normal continuous distribution. Since population distribution is not normal (at least not given specifically), we could expect our confidence interval to be **approximate** only. Population SD σ is given as known which is 3.6 minutes. The sample mean of sample set is 16.2 minutes, thus $x = 16.2$

Summarizing,

$$\begin{aligned} x &= 16.2, n = 120, \sigma = 3.6 \\ 1 - \alpha &= 0.92, \alpha = 0.08, \frac{\alpha}{2} = 0.04 \end{aligned}$$

Since resulting sampling distribution is normal, we could use Z distribution. Remember, we use right tailed Z table here. Recall 1.2.2. Using [this](#) table, we get

$$z_{\frac{\alpha}{2}} = z_{0.04} = 1.75$$

Using 1.19,

$$\begin{aligned} Pr\left(x - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq x + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) &\approx 1 - \alpha \\ Pr\left(16.2 - 1.75 \frac{3.6}{\sqrt{120}} \leq \mu \leq 16.2 + 1.75 \frac{3.6}{\sqrt{120}}\right) &\approx 0.92 \\ Pr\left(16.2 - 0.575 \leq \mu \leq 16.2 + 0.575\right) &\approx 0.92 \\ Pr\left(15.625 \leq \mu \leq 16.775\right) &\approx 0.92 \end{aligned}$$

Thus the 92% confidence intervals for given sample set is [15.625,16.775]

2.2.3 σ Unknown, Population Normal, Low Sample Size

To assess the accuracy of a laboratory scale, a standard weight that is known to weigh 1 gram is repeatedly weighed 4 times. The resulting measurements (in grams) are: 0.95, 1.02, 1.01, 0.98. Assume that the weighings by the scale when the true weight is 1 gram are normally distributed with mean μ . Use these data to compute a 95% confidence interval for μ

Solution:

The population is given as normally distributed with σ unknown. Due to low sample size $n = 4 < 30$, the resultant sampling distribution would be of student's t distribution, than normal, so we need to use that.

Parameters of the sample set:

```
In[22]: x = [0.95, 1.02, 1.01, 0.98]

def get_metrics(x):
    from math import sqrt
    n = len(x) # sample size
    x_bar = sum(x)/n # unbiased sample mean
    var = sum( [(x_i - x_bar)**2 for x_i in x] )/(n-1)
    s = round(sqrt(var),3) # unbiased sample SD
    return n, x_bar, var, s

n,x_bar,_,s = get_metrics(x)
print('n:{} x_bar:{} s:{}'.format(n,x_bar,s))
```

```
n:4 x_bar:0.99 s:0.032
```

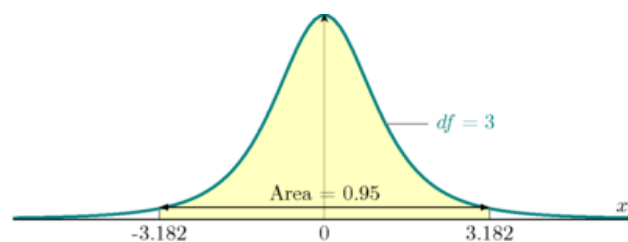
Summarizing,

$$n = 4, \bar{x} = 0.99, s = 0.032, 1 - \alpha = 0.95$$

$$t_{\frac{\alpha}{2},(n-1)} = t_{\frac{0.05}{2},3} = t_{0.025,3}$$

Using [right tailed t table](#), $t_{0.025,3} = 3.182$

If we continued taking sample sets of this size $n = 4$, we would end up getting a sampling distribution that has student's t distribution as below.



Sampling Distribution has t distribution for low sample sizes

Thus, using 1.20,

$$\begin{aligned} Pr\left(x - t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq x + t_{\frac{\alpha}{2},(n-1)} \frac{s}{\sqrt{n}}\right) &= 1 - \alpha \\ Pr\left(0.99 - t_{(0.025,3)} \frac{0.032}{\sqrt{4}} \leq \mu \leq 0.99 + t_{(0.025,3)} \frac{0.032}{\sqrt{4}}\right) &= 0.95 \\ Pr\left(0.99 - 3.182 \frac{0.032}{\sqrt{4}} \leq \mu \leq 0.99 + 3.182 \frac{0.032}{\sqrt{4}}\right) &= 0.95 \end{aligned}$$

```
In[25]: def get_CI(x_bar, zrt, s, n):
        from math import sqrt
        m = zrt*(s/(sqrt(n)))
        return [x_bar-m,x_bar+m]

        t = 3.182
        print(get_CI(x_bar, t, s, n))
```

[0.939088, 1.040912]

∴ the 95% CI in our case are,

$$Pr(0.94 \leq \mu \leq 1.04) = 0.95$$

2.2.4 σ Unknown, Population not Normal, High Sample Size

In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the sample mean and sample standard deviation of number of concurrent users at 100 randomly selected times is 37.7 and 9.2, respectively. Construct a 90% confidence interval for the mean number of concurrent users.

Solution

The measurement at hand is mean number of concurrent users. This is a continuous random variable. Irrespective of population distribution, if sample size is large enough, due to CLT, eventually the sampling distribution formed will be normal. Here $n = 100 > 30$, so we would at least approximately could get good enough CI with 90% confidence level as asked.

Summarizing,

$$n = 100, x = 37.7, s = 9.2$$

$$1 - \alpha = 0.9, \alpha = 0.1, \frac{\alpha}{2} = 0.05$$

This time, we shall use code to find the right tailed z area, ..

```
In[26]: def get_z(cl):
        from scipy import stats
        alpha = round((1 - cl)/2,3)
        return (-1)*(round(stats.norm.ppf(alpha),3)) # right tailing..

        print(get_z(0.90))
```

1.645

Thus, $z_{0.05} = 1.645$ Using 1.21, but also using approximation as we do not know population distribution,

$$\begin{aligned} Pr\left(x - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq x + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right) &\approx 1 - \alpha \\ Pr\left(37.7 - z_{0.05} \frac{9.2}{\sqrt{100}} \leq \mu \leq 37.7 + z_{0.05} \frac{9.2}{\sqrt{100}}\right) &\approx 0.9 \\ Pr\left(37.7 - 1.645 \frac{9.2}{\sqrt{100}} \leq \mu \leq 37.7 + 1.645 \frac{9.2}{\sqrt{100}}\right) &\approx 0.9 \end{aligned}$$

```
In[27]: x, z, s, n = 37.7, 1.645, 9.2, 100
        print(get_CI(x, z, s, n))
```

[36.186600000000006, 39.2134]

Thus the desired 90% CI intervals are [36.2,39.2]

Note: Since the sample size is high, even if t distribution is used, result would be almost same, because at such high sample sizes, t distribution would be almost identical to z distribution.

2.2.5 Difference between two means, Welch's 't' interval

The species, the deinopis and menneus, coexist in eastern Australia. The following summary statistics were obtained on the size, in millimeters, of the prey of the two species. Calculate the 95% confidence interval for the difference in their means.

Adult Dinopis	Adult Menneus
n=10	m=10
$\bar{x} = 10.26mm$	$\bar{y} = 9.02mm$
$s_x^2 = (2.51)^2$	$s_y^2 = (1.90)^2$

Solution

Given:

Let $\bar{X} = N(\mu_{\bar{x}}, \sigma_{\bar{x}}^2)$ be the random variable of sampling distribution for Adult Dinopis. And so is $\bar{Y} = N(\mu_{\bar{y}}, \sigma_{\bar{y}}^2)$ for Adult Menneus. Then we are given one sample set data frame from each species.

$$\bar{x}_1 = 10.26mm, \quad s_{\bar{x}} = 2.51 \text{ mm}, \quad n = 10$$

$$\bar{y}_1 = 9.02mm, \quad s_{\bar{y}} = 1.90 \text{ mm}, \quad m = 10$$

$$1 - \alpha = 0.95, \alpha = 0.05, \frac{\alpha}{2} = 0.025$$

Approach:

Note the σ_x, σ_y are unknown. Also both n, m are small $n < 30, m < 30$. It is totally not needed that $n = m$, but in this case we have that. Recalling 1.25 and 1.26,

$$Pr\left((\bar{X} - \bar{Y}) - t_{(\frac{\alpha}{2}, r)} s_w \leq (\mu_{\bar{x}} - \mu_{\bar{y}}) \leq (\bar{X} - \bar{Y}) + t_{(\frac{\alpha}{2}, r)} s_w\right) \approx 1 - \alpha$$

$$\bar{x}_1 - \bar{y}_1 = 10.26 - 9.02$$

$$s_w = \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}} = \sqrt{\frac{2.51^2}{10} + \frac{1.90^2}{10}}$$

$$r = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{1}{n-1}\left(\frac{s_x^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{s_y^2}{m}\right)^2} = \frac{\left(\frac{2.51^2}{10} + \frac{1.90^2}{10}\right)^2}{\frac{1}{9}\left(\frac{2.51^2}{10}\right)^2 + \frac{1}{9}\left(\frac{1.90^2}{10}\right)^2}$$

```
In[28]: x_1, y_1, s_xbar, s_ybar, n, m = 10.26, 9.02, 2.51, 1.90, 10, 10
```

```
w_1 = round(x_1 - y_1,3)

def get_s_w(s_x, s_y,n,m):
    v_x, v_y = (s_x**2)/n, (s_y**2)/m
    from math import sqrt
    return round(sqrt(v_x + v_y),4)

s_w = get_s_w(s_xbar, s_ybar, n, m)

def get_r(s_x, s_y,n,m):
    v_x, v_y = (s_x**2)/n, (s_y**2)/m
    num = (v_x + v_y)**2
    den_1 = (1/(n-1))*((v_x)**2)
    den_2 = (1/(m-1))*((v_y)**2)
    r = num / (den_1 + den_2)
    from math import modf
    return modf(r)[1]

r = get_r(s_xbar, s_ybar, n, m)

print('x_bar - y_bar:{}'.format(w_1), 's_w:{}'.format(s_w), 'r:{}'.format(r))

# calculate t value
c1 = 0.95
half_alpha = round((1 - c1)/2,3)
from scipy import stats
t = round(stats.t.ppf(1-half_alpha, r),3)

print('t:' + str(t))
```

```
x_bar - y_bar:1.24, s_w:0.9955, r:16.0
t:2.12
```

$$Pr\left((\bar{X} - \bar{Y}) - t_{(\frac{\alpha}{2}, r)} s_w \leq (\mu_{\bar{x}} - \mu_{\bar{y}}) \leq (\bar{X} - \bar{Y}) + t_{(\frac{\alpha}{2}, r)} s_w\right) \approx 1 - \alpha$$

$$Pr\left(1.24 - (2.12)(0.9955) \leq (\mu_{\bar{x}} - \mu_{\bar{y}}) \leq 1.24 + (2.12)(0.9955)\right) \approx 0.95$$

```
In[29]: cilow, cihigh = round((w_1 - t*s_w),4),round((w_1 + t*s_w),4)
print(cilow, cihigh)
```

```
-0.8705 3.3505
```

$$Pr(-0.87 \leq (\mu_{\bar{x}} - \mu_{\bar{y}}) \leq 3.35) \approx 0.95$$

Thus the 95% confidence intervals for the difference of sample means of given problem is $(-0.87, 3.35)$

2.2.6 Difference between two proportions

Duncan is investigating if residents of a city support the construction of a new high school. He's curious about the difference of opinion between residents in the North and South parts of the city. He obtained separate random samples of voters from each region. Here are the results:

Supports Construction?	North	South
Yes	54	77
No	66	63
Total	120	140

Duncan wants to use these results to construct a 90% confidence interval to estimate the difference in the proportion of residents in these regions who support the construction project ($p_S - p_N$). Assume that all of the conditions for inference have been met. Calculate 90% confidence interval based on Duncan's samples

Solution:

Conveniently the sample sizes are high, so we could assume normal approximations for sampling distributions of sample proportions for both North and South parts of the city.

Given:

Let $\frac{Y_S}{n_S} = N\left(p_1, \frac{p_1 q_1}{n_1}\right)$ represent sampling distribution for South. Similarly, $\frac{Y_N}{n_N} = N\left(p_2, \frac{p_2 q_2}{n_2}\right)$ for North.

We have the test statistic as follows.

$$\hat{p}_S = \frac{y_S}{n_S} = \frac{77}{140}, \hat{q}_S = \frac{y_S}{n_S} = 1 - \frac{77}{140}$$

$$\hat{p}_N = \frac{y_N}{n_N} = \frac{54}{120}, \hat{q}_N = 1 - \frac{y_N}{n_N} = 1 - \frac{54}{120}$$

$$1 - \alpha = 0.90, \alpha = 0.1, \frac{\alpha}{2} = 0.05$$

```
In[12]: t_s = [77/140, 1-(77/140), 54/120, 1-(54/120)]
t_s = ['%0.3f' % e for e in t_s]
t_s = [float(i) for i in t_s]
[p_s, q_s, p_n, q_n] = t_s
print(p_s, q_s, p_n, q_n)
```

0.55 0.45 0.45 0.55

$\therefore \hat{p}_S = 0.55, \hat{q}_S = 0.45, \hat{p}_N = 0.45, \hat{q}_N = 0.55$ Recalling 1.27, we need to find,

$$Pr\left(-z_{\frac{\alpha}{2}} \leq \frac{(\hat{p}_S - \hat{p}_N) - (p_S - p_N)}{\sqrt{\frac{\hat{p}_S \hat{q}_S}{n_S} + \frac{\hat{p}_N \hat{q}_N}{n_N}}} \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha = 0.90$$

```
In[16]: diff = round(p_s - p_n,3)

n_s, n_n = 140,120
from math import sqrt
w_sd = round(sqrt((p_s*q_s/n_s) + (p_n*q_n/n_n)),3)
```

```
# get Z
c1 = 0.90
from scipy import stats
alpha = 1 - c1
z = (-1)*round(stats.norm.ppf(alpha/2),3)

print(diff, w_sd, z)
```

0.1 0.062 1.645

Substituting, we get,

$$\Pr\left(-1.645 \leq \frac{0.1 - (p_S - p_N)}{0.062} \leq 1.645\right) \approx 0.90$$

$$\Pr\left((-1.645)0.062 \leq 0.1 - (p_S - p_N) \leq (1.645)0.062\right) \approx 0.90$$

$$\Pr\left(0.1 - (1.645)0.062 \leq (p_S - p_N) \leq 0.1 + (1.645)0.062\right) \approx 0.90$$

```
In[18]: cilow, cihigh = round(diff - z*w_sd,3), round(diff + z*w_sd,3)
print(cilow, cihigh)
```

-0.002 0.202

Thus the 90% CI intervals for the difference between proportions are $(-0.002, 0.202)$. That is,

$$\Pr\left(-0.002 \leq (p_S - p_N) \leq 0.202\right) \approx 0.90$$

2.3 Useful Snippets

2.3.1 Python

Get t score

Could be useful, when you have significance level α and degrees of freedom $df = n - 1$, and have to calculate corresponding t score

```
In[30]: def get_t(c1, n):
        from scipy import stats
        half_alpha = round((1 - c1)/2,3)
        return round(stats.t.ppf(1-half_alpha, n-1),3)

c1 = 0.95 # confidence level
n = 4    # sample size
print(get_t(c1, n))
```

3.182

Get Z score

Could be useful, when you have significance level α and have to calculate corresponding Z score. Remember to always check if you need left tailed area or right tailed.

```
In[31]: def get_z(c1):
        #NOTE:returns right tailed area as that is mostly used in CI
        from scipy import stats
        alpha = round((1 - c1)/2,3)
        return (-1)*round(stats.norm.ppf(alpha),3) # right tailing..

c1 = 0.95
print(get_z(c1))
```

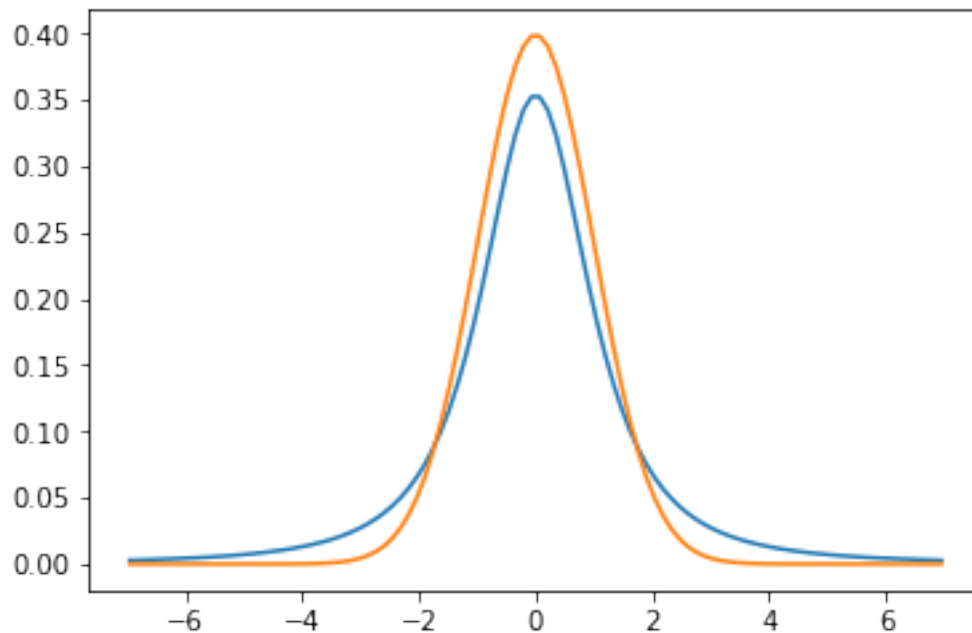
1.96

Z and T distribution

Plotting a z and t distribution.

```
In[32]: %matplotlib inline
        from scipy.stats import t, norm
        import numpy as np
        import matplotlib.pyplot as plt

n = 3
df = n-1
fig,ax = plt.subplots(1,1)
x = np.linspace(t.ppf(0.01,df), t.ppf(0.99,df),100)
ax.plot(x, t.pdf(x,df), color='C0') # blue is t distribution
ax.plot(x, norm.pdf(x), color='C1') # red
plt.show()
```



2.3.2 Tikz in Ipython

Some parts of this book including this section are created using ipython notebooks and thus few figures which needed to be constructed via tikz needed an extension. Below figures are created via

tikz by using an ipython extension called `tikzmagic`, so the format is slightly different for preamble. However, for tikz users, the essence could be easily captured.

For first time usage (or after reset and clear of notebook), always load tikz as below.

```
%load_ext tikzmagic
```

Also note, preamble is placed in a separate code cell above, because ipython needs magic commands to start as first line in cells. Here, tikz execution needs a magic command in subsequent cell.

Z distribution:

```
In[33]: preamble = '''
        \pgfmathdeclarefunction{gauss}{3}{%
        \pgfmathparse{1/(#3*sqrt(2*pi))*exp(-((#1-#2)^2)/(2*#3^2))}%
        }
        '''
```

```
In[34]: %%tikz -p pgfplots -x $preamble
        % had to be this size to have a normal size in latex
        \begin{axis}[
            no markers,
            domain=0:6,
            samples=100,
            ymin=0,
            axis lines*=left,
            xlabel=$x$,
            ylabel=$f(x)$,
            height=5cm,
            width=12cm,
            xtick=\empty,
            ytick=\empty,
            enlargelimits=false,
            clip=false,
            axis on top,
            grid = major,
            axis lines = middle
        ]

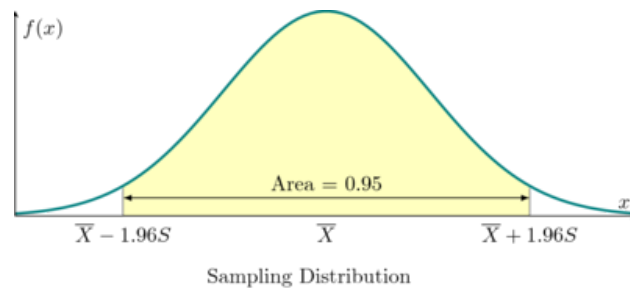
        \def\mean{3}
        \def\sd{1}
        \def\cilow{\mean - 1.96*\sd}
        \def\cihigh{\mean + 1.96*\sd}
        \addplot [draw=none, fill=yellow!25, domain=\cilow:\cihigh] {gauss(x, \mean, \sd)}
        \closedcycle;
        \addplot [very thick,cyan!50!black] {gauss(x, 3, 1)};

        \pgfmathsetmacro\valueA{gauss(1,\mean,\sd)}
        \draw [gray] (axis cs:\cilow,0) -- (axis cs:\cilow,\valueA) (axis cs:\cihigh,0) --
        (axis cs:\cihigh,\valueA);
        \draw [yshift=0.3cm, latex-latex](axis cs:\cilow, 0) -- node [above] {Area = $0.95$}
        (axis cs:\cihigh, 0);

        \node[below] at (axis cs:\cilow, 0) {$\overline{X} - 1.96S$};
        \node[below] at (axis cs:\mean, 0) {$\overline{X}$};
        \node[below] at (axis cs:\cihigh, 0) {$\overline{X} + 1.96S$};

        \node[below=0.75cm,text width=4cm] at (axis cs:\mean, 0){Sampling Distribution};

        \end{axis}
```

***t* distribution:**

```
In[35]: preamble='''
        \pgfmathdeclarefunction{gamma}{1}{%
            \pgfmathparse{2.506628274631*sqrt(1/#1)+ 0.20888568*(1/#1)^(1.5)+
            0.00870357*(1/#1)^(2.5)- (174.2106599*(1/#1)^(3.5))/25920-
            (715.6423511*(1/#1)^(4.5))/1244160)*exp((-ln(1/#1)-1)*#1)%
        }

        \pgfmathdeclarefunction{student}{2}{%
            \pgfmathparse{gamma((#2+1)/2.)/(sqrt(#2*pi) *gamma(#2/2.))
            *((1+(#1*#1)/#2)^(-#2+1)/2.)}%
        }
        '''
```

```
In[36]: %%tikz -p pgfplots -x $preamble
        \begin{axis}[
            no markers,
            domain=-6:6,
            samples=100,
            ymin=0,
            axis lines*=left,
            xlabel=$x$,
            height=5cm,
            width=12cm,
            xtick=\empty,
            ytick=\empty,
            enlargelimits=false,
            clip=false,
            axis on top,
            grid = major,
            axis lines = middle,
            y axis line style={draw opacity=0.25}
        ]
            \def\mean{0}
            \def\sd{1}
            \def\df{3}
            \def\cilow{-3.182}
            \def\cihigh{3.182}

            \addplot [draw=none, fill=yellow!25, domain=\cilow:\cihigh] {student(x, \df)}
            \closedcycle;
            \addplot [very thick,cyan!50!black] {student(x, \df)} node [pos=0.6, anchor=mid
            west, xshift=2em, append after command={{\tikzlastnode.west} edge [thin, gray]
            +(-2em,0)}] {$df=3$};

            %https://tex.stackexchange.com/questions/453059/pgfmathsetmacro-creates-dimensions-
            %too-large-for-t-distribution/453062
            \addplot [ycomb, gray, no markers, samples at={\cilow, \cihigh}] {student(x, \df)};
            \draw [yshift=0.2cm, latex-latex](axis cs:\cilow, 0) -- node [above] {Area = $0.95$}
            (axis cs:\cihigh, 0);
```

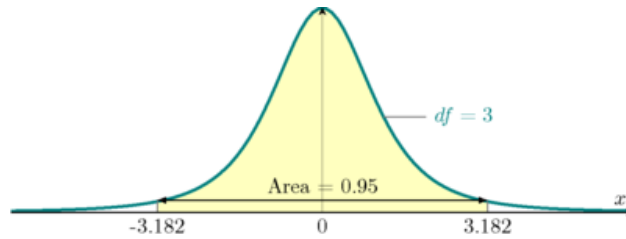
```

\node[below] at (axis cs:\cilow, 0) {\cilow};
\node[below] at (axis cs:\mean, 0) {0};
\node[below] at (axis cs:\cihigh, 0) {\cihigh};

\node[below=0.75cm,align=center, text width=10cm] at (axis cs:\mean, 0){Sampling
Distribution has  $t$  distribution for low sample sizes};

\end{axis}

```



Sampling Distribution has t distribution for low sample sizes

Chapter 3

Appendix

3.0.1 Difference between two Random Variables

Suppose we have two populations described by random variables $X(\mu_x, \sigma_x^2)$ and $Y(\mu_y, \sigma_y^2)$. We are interested in the distribution of their differences $W = X - Y$. What would be μ_w, σ_w^2 ? We could solve that using Expectations. This is true for any distributions X and Y have, as long as they are independent to each other or $X \neq Y$.

$$\mu_w = E[W] = E[X - Y] = E[X] - E[Y] = \mu_x - \mu_y \quad (3.1)$$

$$\begin{aligned} \sigma_w^2 &= Var[W] = E[W^2] - [E[W]]^2 \\ &= E[(X - Y)^2] - [E[X - Y]]^2 \\ &= E[X^2 + Y^2 - 2XY] - (E[X] - E[Y])^2 \\ &= E[X^2] + E[Y^2] - E[2XY] - \left\{ (E[X])^2 + (E[Y])^2 - 2E[X]E[Y] \right\} \\ &= E[X^2] + E[Y^2] - \underline{2E[X]E[Y]} - (E[X])^2 - (E[Y])^2 + \underline{2E[X]E[Y]} \\ &= \{E[X^2] - (E[X])^2\} + \{E[Y^2] - (E[Y])^2\} \\ &= Var[X] + Var[Y] \end{aligned}$$

$$\therefore \sigma_w^2 = \sigma_x^2 + \sigma_y^2 \quad (3.2)$$

Bibliography

- [1] L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. 16, 2001. URL <https://projecteuclid.org/euclid.ss/1009213286>.
- [2] Robert, Elliot, and Dale. *Probability and Statistical Inference*. Pearson, 9th edition, 2015. URL <http://www.nylxs.com/docs/thesis/sources/Probability%20and%20Statistical%20Inference%209ed%20%5B2015%5D.pdf>.