CHAPTER 23

23 Confidence Intervals for the mean

Instead of offering a a point estimate,
it is often more realistic/accurate to provide a range for an estimate.

“It is going to rain tomorrow at 3:54.” versus
“It is going to rain tomorrow sometime between 3:00 and 5:00”

“We believe the missing boat is at 33.98 Lat and -117.37 Long.” versus
”We believe the missing boat is within a 1-mile radius of 33.98 Lat and -117.37 Long.”

”Presidential candidate X is going to win.” versus
“Presidential candidate X is leading the polls 56% to 49% with a margin of error of 5% with a 95% confidence interval.”

When we sample a population, we know that the mean of the sample is close to the mean of the population; similarly with the variance.

That is to say: $\bar{x}$ and s^2 (aka sample statistics) are close to
μ and σ^2 (aka population parameters). Estimates vs. facts.

However, we would like to judge our estimates. What does 'close' mean?

This is better answered with an interval as opposed to a point estimate.
Along with the interval of values we will also provide an associated level of confidence for the interval.

Example:
"The final exam scores will have an average of:
89 plus/minus 3 points with 95% confidence."

Statistical Inference

Goal: to extract properties of an underlying population by analyzing the data of a sample. Therefore, we need answers to questions about statistics and parameters:

1. How well does $\bar{x}$ approximate μ?

2. Is the sample proportion $\hat{p}$ a good approximation of the population proportion p?

3. Is there a statistically significant difference between the mean of two samples?
   If the answer is yes, how sure are we?

4. How much data do we need in order to be confident in our conclusion?

This is a 95% Confidence Interval for μ:

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \leq \quad \mu \quad \leq \quad \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

The Central Limit Theorem tells us that as the sample size n increases,
the sample mean of X is close to normally distributed,
with expected value μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{X} \sim N\left(\mu, \sigma^2/_n\right)$$

Standardizing the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable.

$$Z = \frac{\bar{X}-\mu}{\sigma/_{\sqrt{n}}} \sim N(0,1)$$

This has $\bar{X}$ ‘living’ in the $N(0,1)$ neighborhood

So, $\bar{X}$ is normally distributed, and after standardizing we find 95% of the area under the standard normal to be between -1.96 and $+1.96$.

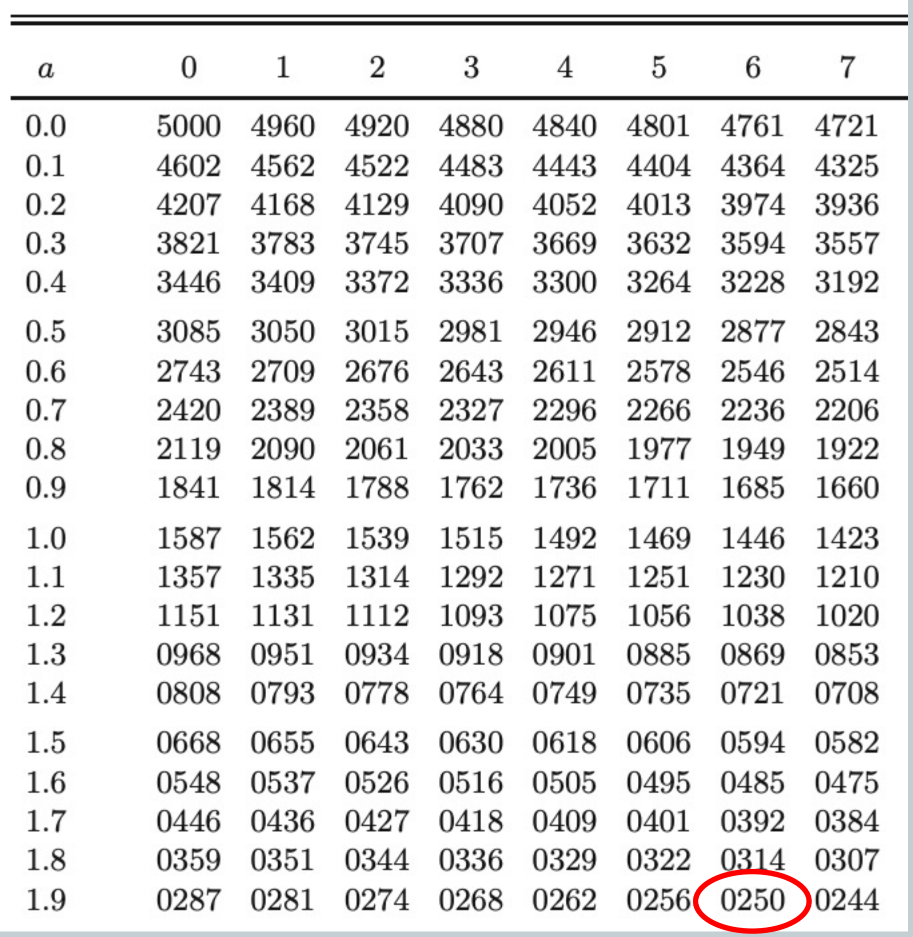
$$P(-1.96 \le Z \le 1.96) = 0.95$$

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

$$P\left(-1.96 \cdot \frac{\sigma}{\sqrt{n}} \le \bar{X} - \mu \le 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(-\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \le \quad \mu \quad \le \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.95$$



a	0	1	2	3	4	5	6	7
0.0	5000	4960	4920	4880	4840	4801	4761	4721
0.1	4602	4562	4522	4483	4443	4404	4364	4325
0.2	4207	4168	4129	4090	4052	4013	3974	3936
0.3	3821	3783	3745	3707	3669	3632	3594	3557
0.4	3446	3409	3372	3336	3300	3264	3228	3192
0.5	3085	3050	3015	2981	2946	2912	2877	2843
0.6	2743	2709	2676	2643	2611	2578	2546	2514
0.7	2420	2389	2358	2327	2296	2266	2236	2206
0.8	2119	2090	2061	2033	2005	1977	1949	1922
0.9	1841	1814	1788	1762	1736	1711	1685	1660
1.0	1587	1562	1539	1515	1492	1469	1446	1423
1.1	1357	1335	1314	1292	1271	1251	1230	1210
1.2	1151	1131	1112	1093	1075	1056	1038	1020
1.3	0968	0951	0934	0918	0901	0885	0869	0853
1.4	0808	0793	0778	0764	0749	0735	0721	0708
1.5	0668	0655	0643	0630	0618	0606	0594	0582
1.6	0548	0537	0526	0516	0505	0495	0485	0475
1.7	0446	0436	0427	0418	0409	0401	0392	0384
1.8	0359	0351	0344	0336	0329	0322	0314	0307
1.9	0287	0281	0274	0268	0262	0256	0250	0244

The 95% Confidence Interval is given by the values of X that satisfy the equation:

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \leq \quad \mu \quad \leq \quad \bar{X} + \ 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

How large does the sample need to be if

1] The population is normally distributed?

2] The population is not normally distributed?

3] Which things are random? Which are fixed?

The 95% Confidence Interval is given by the values of X that satisfy the equation:

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \quad \leq \quad \mu \quad \leq \quad \bar{X} + \ 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

How large does the sample need to be if

1] The population is normally distributed?

$n \geq 1$

2] The population is not normally distributed?

$n \geq 30$ So say some statisticians…

3] Which things are random? Which are fixed?

$\bar{X}$ is random.

$\sigma, n,$ and 1.96 are fixed.

Although the 1.96 is specifically associated with a 95% level of confidence.

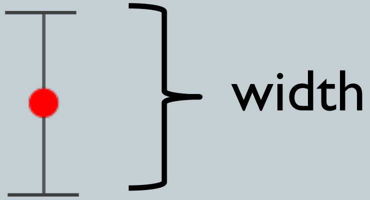
Suppose you want to discover the mean weight of a buffalo from a herd.

You collect a sample 35 buffaloes and find their mean weight and compute a 95% CI.
Then you return the 35 buffalo to the herd.
You collect another sample of 35 buffaloes and find their mean weight and compute a 95% CI.
Then you return the 35 buffalo to the herd.
You collect another sample of 35 buffaloes and find their mean weight and compute a 95% CI.
Then you return the 35 buffalo to the herd.
etc. etc. etc.

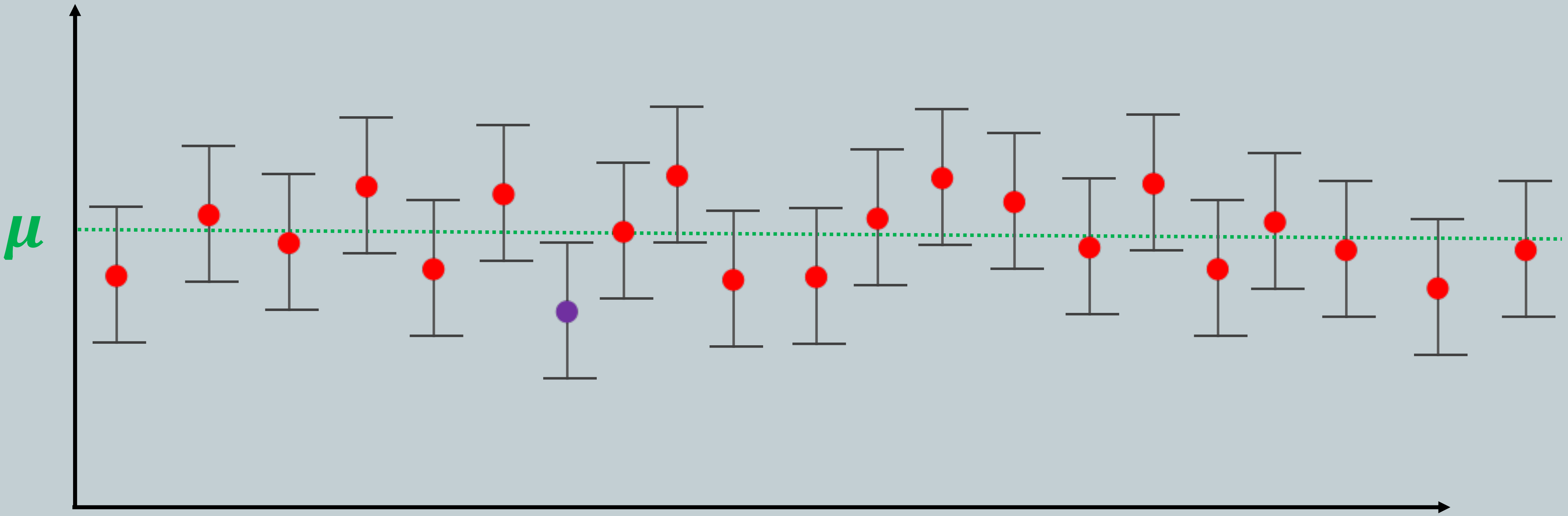
Ultimately, let's say you collect 20 random samples of 35 buffaloes.
How many of these 20 intervals do you expect to contain the true population mean μ ?

That is to say, how many of your CI's should contain the herds true average buffalo weight?

All 95%-CI's for a particular population with the same sample size, n, and known σ are the same width.
The only thing that changes is the center, $\bar{x}$.



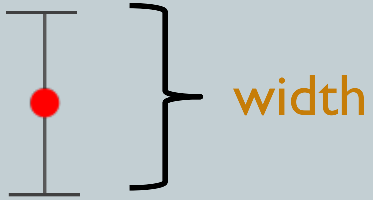
With 20 random samples we expect $0.95 \cdot 20 = 19$ of them to contain the true mean, μ.



The 95%-CI is centered at $\bar{x}$ and extends $1.96 \cdot \frac{\sigma}{\sqrt{n}}$ to each side of $\bar{x}$.

The CI's width is $2 \cdot 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ which is not random.

Only the location of the interval's midpoint $\bar{x}$ is random.



We often write the CI $\bar{x} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

as $\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$

Why are we talking about 95% confidence intervals?
specifically, why 95% ? Why not 94% or 96% or 50% or…

Why is 95% a standard? (journal articles)
Who wants to make a $100 bet?

What if I need more confidence; say 99% ? (medical studies, tolerance)

What if less accuracy is required; 75%? (smaller study sample, time, $$)

- For 95% CI :

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

In general:

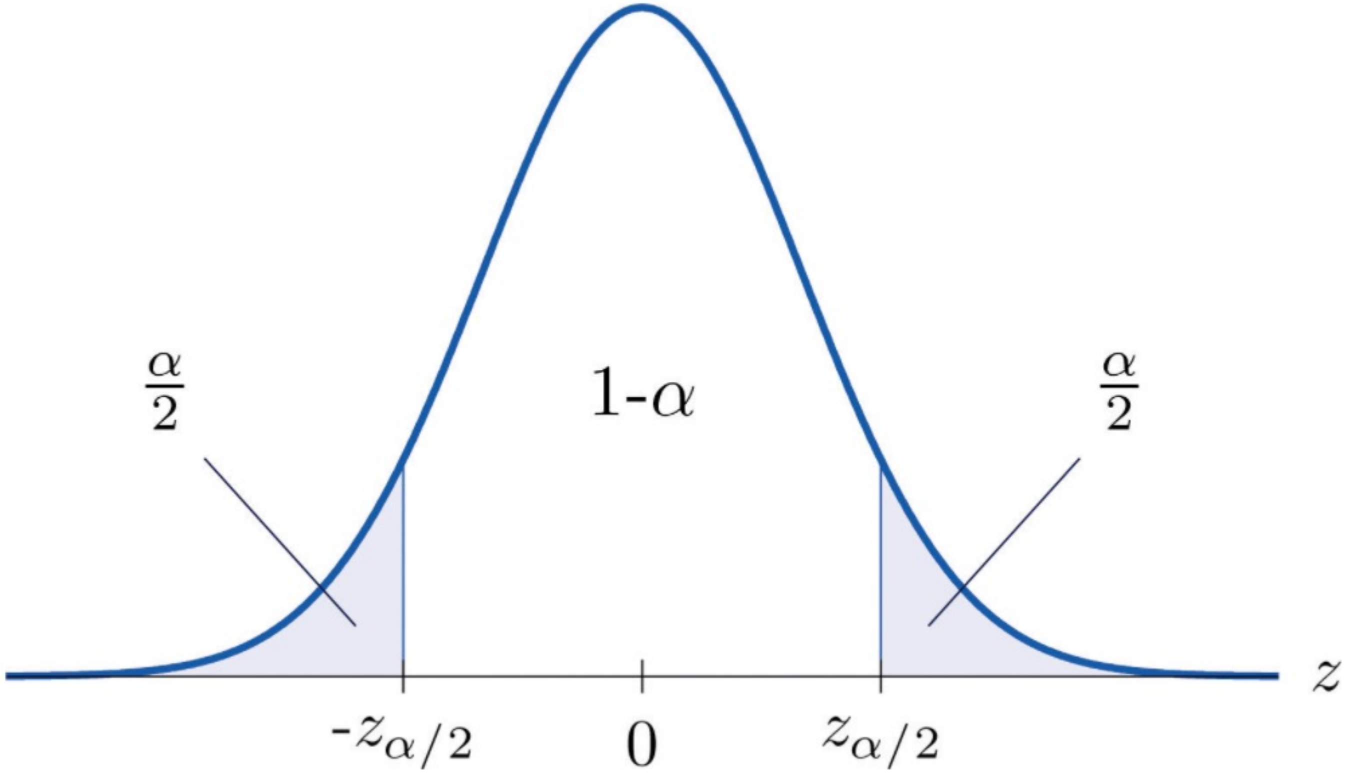
for $100(1-\alpha)^{th}$ CI

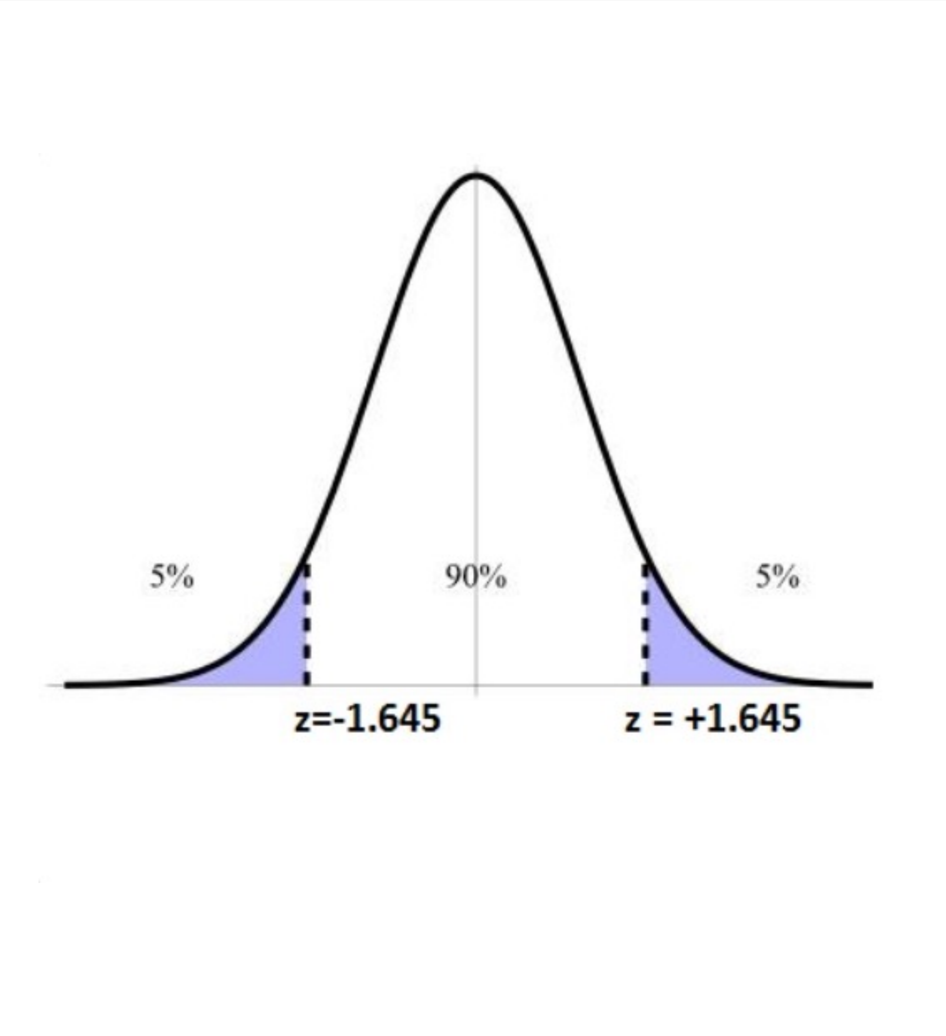
$$P(\_\_\_\_ \le Z \le \_\_\_\_) = 1-\alpha$$

- A $100(1-\alpha)^{th}$ CI :

$$\left[\ \bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\ ,\ \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\ \right]$$

- Just name the desired α





Suppose we are after a 90% CI

Then $\alpha = 0.1$

$$\alpha/_2 = {}^{0.1}/_2 = 0.05$$

90% CI : $[\bar{x} - z_{.05} \cdot \frac{\sigma}{\sqrt{n}} \quad , \bar{x} + z_{.05} \cdot \frac{\sigma}{\sqrt{n}}]$

How does one find $z_{.05}$?

Either by a table, or

```
stats.norm.ppf(1-0.05)
```

$z_{.05} = 1.645$

We want to know the average amount of time a US resident spends with family. A random sample of 1000 residents reveals an average of 3.6 hours per day, and we have a standard deviation of 2 hours per day.

Find a 90% CI for the true population mean, μ.

ANS:

The above information provides: $\bar{x} = 3.6$, $\sigma = 2$, $\alpha = 0.1$, therefore:

$$\left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad , \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = \left[3.6 - z_{.05} \cdot \frac{2}{\sqrt{1000}} \quad , 3.6 + z_{.05} \cdot \frac{2}{\sqrt{1000}}\right]$$

$z_{.05} =$ `stats.norm.ppf(.95)` $= 1.645$ or $\Phi(z_{\alpha/2}) = 0.95$

3.6 ± 0.104 or a 90% CI for the population mean is $[3.496 , 3.704]$.

"My estimate of the population mean is 3.6 with a 90% CI of [3.496 , 3.704]."

In the previous example we found a 90% CI.
What if we wanted a 95% CI instead?

We believe $\mu = 3.6$ with CI $[3.496\,,3.704]$ versus
We believe $\mu = 3.6$ with CI $[3.476\,,3.724]$.

Which is the 90% CI and which is the 95% CI ?

The only difference in the 95% CI is $z_{\alpha/2}$: 1.645 versus 1.96

$$\left[3.6 - z_{.05} \cdot \frac{2}{\sqrt{1000}} \quad ,3.6 + z_{.05} \cdot \frac{2}{\sqrt{1000}}\right]$$

$$\left[3.6 - z_{.025} \cdot \frac{2}{\sqrt{1000}} \quad ,3.6 + z_{.025} \cdot \frac{2}{\sqrt{1000}}\right]$$

So, the more ‘confidence’ you seek, the wider the interval becomes.
Technically, I could say with 100% confidence that the actual exam average is going to be between 0 and 100. But that isn’t the most useful estimate.

What if I want the best of both worlds? High confidence **and** a small interval?
Suppose I seek a 95% CI with width at most 0.1 ? How can **this** be accomplished?

$$\left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad , \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

How can we create a small interval with a high confidence level? For instance…

Suppose I seek a 95% CI with width at most 0.1 ? How can this be accomplished?

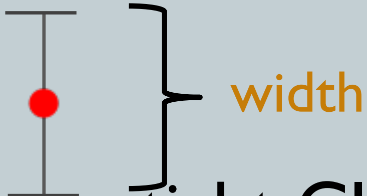
$$\left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad , \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Increase n, the sample size. Makes sense…

Specifically, if $\sigma = 2$ as in our family-time example, then at 95% CI we have:

$2 \cdot 1.96 \cdot \frac{2}{\sqrt{n}} \leq 0.1$ (2 creates the width, double…)

$2 \cdot 19.6 \cdot 2 \leq \sqrt{n}$



$6146.57 \leq n$, therefore $n \geq 6147$. Large sample, time, money,… tight CI.

Take a close look at our Confidence Intervals:

$$\left[\bar{x} - z_{\alpha/_2} \cdot \frac{\sigma}{\sqrt{n}} \quad , \bar{x} + z_{\alpha/_2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

How often do we actually know the population σ ?

Almost never. So, what do we do?

If n is sufficiently large ($n \geq 30$), we use the sample variance.

$$\left[\bar{x} - z_{\alpha/_2} \cdot \frac{s}{\sqrt{n}} \quad , \bar{x} + z_{\alpha/_2} \cdot \frac{s}{\sqrt{n}}\right]$$

numpy: `np.var(x, ddof=1)` implies $\frac{1}{n-1}\sum\ldots$ (or `np.std`)
pandas: `dataframe_name.var()` automatically adjusted to sample statistic.

https://numpy.org/doc/stable/reference/generated/numpy.std.html

Proportions

Finding CI’s for other things besides the mean

Suppose we want to know the proportion of female buffaloes from a large herd. Perhaps the proportion is 40% or 65% or… who knows?

Take a sample from the herd. We will mark ‘female’ as success.

X is the number of successes in the sample of size n.
Then X can be modeled as a Binomial Random Variable
with $E[X] = np$ and $Var(X) = np(1-p)$.
i.e. $n = 40$ buffalo sample, $p = 0.40$ implies $E[X] = np = 40 \cdot 0.4 = 16$

Of course, p generally isn’t known, so we estimate with $\hat{p} = \frac{X}{n}$.
$X = 16$ female buffalo out of the sample size of $n = 40$, implies $\hat{p} = 0.4$

Proportions

The estimator for p is $\hat{p} = \frac{X}{n}$.

The estimator, $\hat{p}$, is approximately normally distributed $\hat{p} \sim N\left(p\ ,\ \frac{p(1-p)}{n}\right)$

i.e., $E[\hat{p}] = p$ and $Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{p(1-p)}{n}$ why?

Because $Var\left(\frac{X}{n}\right) = Var\left(\frac{1}{n} \cdot X\right) = \frac{1}{n^2} Var(X) = \frac{1}{n^2} \cdot \left(np(1-p)\right) = \frac{p(1-p)}{n}$

$100(1-\alpha)\%$ Confidence Intervals for $\hat{p}$ *is* $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

Where $\sqrt{\frac{p(1-p)}{n}}$ is referred to as the standard error of $\hat{p}$.

And if we don't know p, then $Var(\hat{p}) \approx \frac{\hat{p}(1-\hat{p})}{n}$

The standard error for $\hat{p}$ is given by: $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Confidence levels for Proportions

The EPA considers indoor radon levels of about 4 picocuries/liter (pCi/L) of air to be high enough to warrant amelioration efforts.

Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels of about 4 pCi/L.

Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

Confidence levels for Proportions

Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

Ans:

$n = 200 \qquad x = 127 \qquad \hat{p} = \frac{127}{200} = 0.635$

$99\%\ CI$ means $\alpha = 0.01,$ so `stats.norm.ppf(1-0.005) = 2.5758`

CI: $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ which we will estimate using $\hat{p}$.

$$0.635 \pm 2.57 \cdot \sqrt{\frac{0.635(1-0.635)}{200}} = [0.548, 0.722]$$

Suppose we are estimating the proportion of people in a population with high blood pressure (HBP). Subjects are diagnosed with, or without HBP.

When the outcome of interest is dichotomous, then the investigator defines one outcome as success and the other as failure.

Let the sample size be n and let X denote the number of successes, (‘success’ being a diagnosis with HBP).

If we sample the population, then $\hat{p} = \frac{X}{n}$

A confidence interval can be calculated: $\hat{p} \pm z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ but only if there are more than 5 successes and more than 5 failures.

A Boulder health study on HBP consisted of 1219 participants being treated (i.e. they have HBP) and 2313 who were not on treatment.

So, if we call treatment ”success”, then $X = 1219$ and $n = 3532$.

We seek a 95% CI for the true population proportion of people with HBP.

Therefore, $\hat{p} = \frac{X}{n} = \frac{1219}{3532} = 0.345$, this is our point estimate, with $Z_{0.975} = 1.96$

$$\hat{p} \pm Z_{0.975} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.345 \ \pm 0.016$$

So, the 95% CI for p is $(0.329, 0.361)$.

We can be 95% confident that the true proportion of persons with HBP is between 32.9% and 36.1%.

Or, we are 100% confident that 95% of CI’s constructed in this manner will contain the true population proportion of those with HBP, and this is one such CI.

Algorithm To calculate a CI for a population proportion:

1] Determine the confidence level and find the appropriate z-value.

2]Find the sample proportion, $\hat{p}$, by dividing the number of people in the sample having the characteristic of interest by the sample size, n.

3] Multiply $\hat{p}(1-\hat{p})$ and then divide that amount by n, and then take the square root.

4] Multiply that answer by z. This gives you the standard error.

5] Take $\hat{p}$ plus or minus the margin of error to obtain the CI.

$100(1-\alpha)\%$ Confidence Intervals for $\hat{p}$ *is* $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Confidence intervals for proportions:

http://digitalfirst.bfwpub.com/stats_applet/stats_applet_20_ciprop.html

Next time: Two-Sample confidence intervals