Confidence Interval for a Mean

- Given a sample of <u>*n*</u> number values drawn from a random variable.
- $X_1, \ldots X_n \sim N(\mu, \sigma^2)$
- Assuming we know the variance, σ^2, we can calculate the confidence interval of the (population) mean, μ, using the equation of the confident interval.

- $[U, V] = \left[\bar{X} - \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \bar{X} + \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$

Lesson Summary

We have two factual statements, one analyst definition, and two additional factual conditions.

1. First, we have observed *n* (e.g., 100) values and know the variance. We can denote them X_1, X_2, … X_n (**Statement 1**).
2. Secondly, the definition of Equation 1 (**Statement 2**).
3. The analyst definition is the probability value for α. Let’s use the traditional value, 0.05, for convenience.
4. The first additional fact is Lemma 2.3, “the sample mean of an IID variable is distributed normally with the mean, μ and the standard deviation, named standard error, σ^2/n. That is, $\bar{X} \sim N(\mu, \sigma^2/n)$”
5. The second additional fact is $q_{\alpha/2}$ is the (1 – α/2)-quantile of the **standard** normal distribution.

When we accept these five points, we can then derive the conclusion, “**then [U, V] is a (1- α)-confidence interval for the mean μ,**” and understand the process in the Proof.

Symbols and Words in Detail

$$
\begin{aligned}
&P(\mu < U) \\
&= P\left(\mu < \bar{X} - \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right) \\
&= P\left(\mu - \bar{X} < -\frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X} - \mu > \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right) \\
&= P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > q_{\frac{\alpha}{2}}\right) \\
&= 1 - P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le q_{\alpha/2}\right)
\end{aligned}
$$

- From point 5 in summary (Slide 2) and using α has a value of 0.05 as an example, we accept that $q_{\alpha/2}$ =1.96, i.e., 97.5 % quantile. The 1.96 was derived from the standard normal distribution.
- Illustrated in Figures 1 and 2 (see Slide 3), the blue area occupies 97.5% of the total area. The unshaded area in Figure 2 occupies 95% of the total area under the curve. The two red areas in Figure 2 show the two sides of the tail. Each occupies 2.5% (i.e., $\alpha/2$).

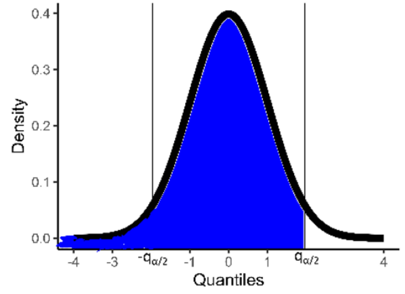


Figure 1. Standard normal distribution 1.

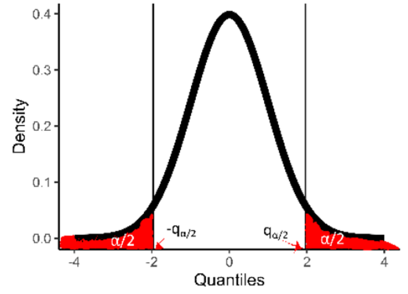


Figure 2.Standard normal distribution 2.

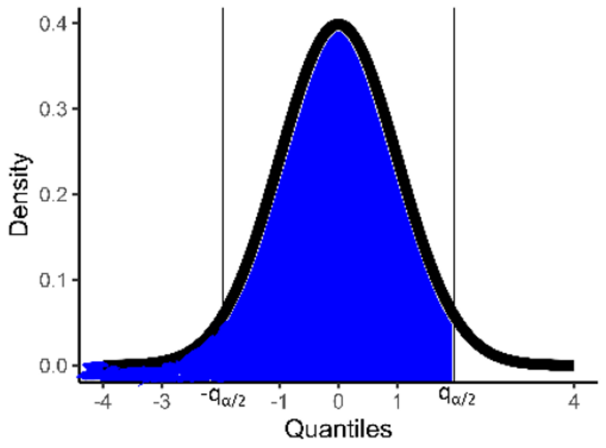


Figure 1. Standard normal distribution 1.

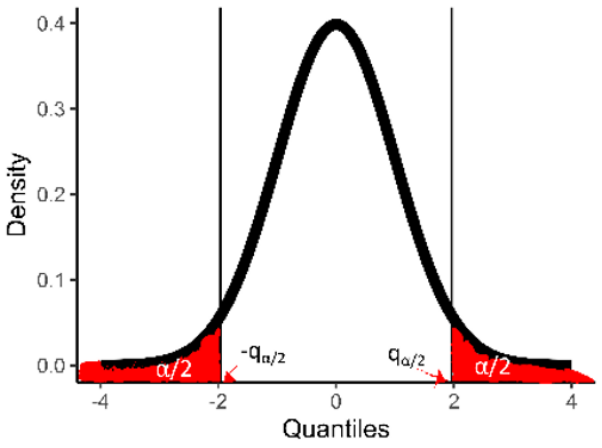


Figure 2.Standard normal distribution 2.

- Therefore, $P(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > q_{\alpha/2})$ is to ask what is the probability that the standardised score of the sample mean greater than 1.96 ($q_{\alpha/2}$). This is the un-shaded area in Figure 1, which is shown as the red area in the right tail in Figure 2. Because the normal distribution is symmetric, the left-side red area has the same probability as the right-side red area.

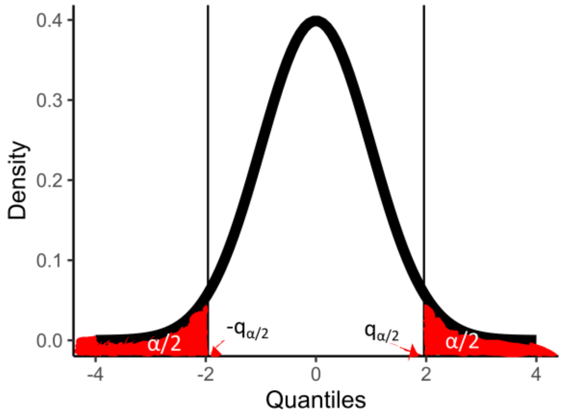
- Writing the process in mathematical symbols, we can derive the last part of the question.

$$P(\mu < U) = 1 - P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq q_{\alpha/2}\right)$$

- Point 4, in summary, gives the lemma, "The sample mean of an IID variable is distributed normally with the mean, μ and the standard error, σ^2/n."

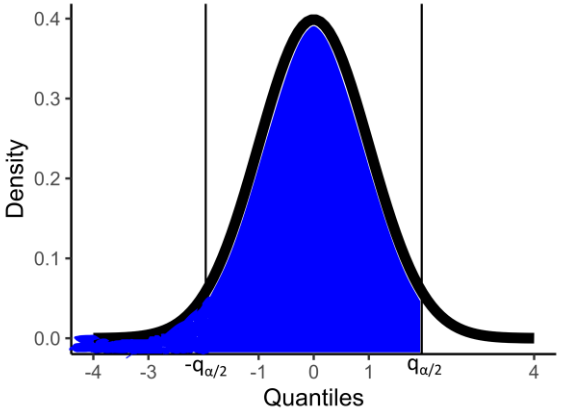
$$\frac{\bar{X} - \mu}{\sigma^2/n} \sim N(0, 1)$$

$\frac{\bar{X}-\mu}{\sigma^2/n} \sim N(0, 1)$; **Equation 3**



- First, we note that the left part of Equation 3 is the standardised score of the sample mean in the distribution, *N(μ, σ²/n)*. This is why we can state, "From Lemma …, thus we have".
- Finally, we can find the probability of (μ<U) using the right-hand side of Equation 3. Using the standard normal distribution, N(0, 1).
- This is illustrated as the blue area in **Figure 1**. $P(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq q_{\alpha/2})$ = the blue area = 1 – α/2. Therefore, P(μ<U)=α/2.
- Going over the same reasoning and process and replace μ < U with μ > V, we can derive, P(μ>V)=α/2 . The final concluding part of the proof is then the arithmetic step of deriving P(μ∈[U,V]), i.e., the unshaded area in **Figure 2**.

$$
\begin{aligned}
&P(\mu < U)\\
&= P\left(\mu < \bar{X} - \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)\\
&= P\left(\mu - \bar{X} < -\frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)\\
&= P\left(\bar{X} - \mu > \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)\\
&= P\left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > q_{\frac{\alpha}{2}}\right)\\
&= 1 - P(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq q_{\alpha/2})
\end{aligned}
$$



Application

- We observed 100 values and dubbed them respective, X_1, X_2, … X_{100}. They may look like this (Statement 1):
- For instance, opinion polls, medical studies, and quality control in manufacturing

```
 [1]  1.612 -1.837  3.128  0.728  3.602  3.724  1.690  5.180  4.834  0.938
[11]  2.499  0.937  7.003 -0.949  7.740  2.723  6.024  2.279  2.107  3.015
[21]  4.313 -1.282  1.530  5.535  0.308  0.171  1.740  4.358  1.321  8.919
[31]  4.289 -3.307 -3.687 -1.125  2.582 -5.288  6.133 -8.514  7.728  2.460
[41] -3.915 -1.096  2.675  5.080 -3.102  1.904  2.239 -6.431  0.362  2.567
[51]  6.263 -4.190  9.525  5.652  0.107  1.082 -1.281  1.772  4.360  7.859
[61]  3.461  3.811  1.139  1.073  3.401  3.423 -2.793  2.963  3.013  1.791
[71] -5.426  2.929  0.722 -2.826  0.951  2.257 -1.374 -1.407  3.386  4.292
[81] -2.731  0.228  6.184  5.627 -2.136  3.404  0.602  0.015  1.325 -3.051
[91] -0.588  9.898  7.974 -0.401 -0.308  2.751  1.049 -2.152  2.754  0.174
```

- We can run a simulation study using a programming language like R or Python.

```
num_samples <- 100
mean <- 2      # Unknown
variance <- 4 # Known
std_dev <- sqrt(variance)
x <- rnorm(num_samples, mean, std_dev)
mean(x)
# 1.72
```

```
num_samples = 100
mean = 2                        # Unknown
variance = 4                    # Known
std_dev = variance ** 0.5
x = np.random.normal(mean, std_dev, num_samples)
xmean = np.mean(x)
# 1.72
```

```
mean(x) - qnorm(0.975) * sqrt(std_dev/n)
mean(x) + qnorm(0.975) * sqrt(std_dev/n)
```

```
U = xmean - norm.ppf(0.975) * np.sqrt(std_dev/num_samples)
V = xmean + norm.ppf(0.975) * np.sqrt(std_dev/num_samples)
```

$$\left[\bar{X} - \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \bar{X} + \frac{q_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right]$$

- In this case, the mean, μ, is unknown, and the variance, 4, is known, and we believe these 100 values are a draw from a **normal** (population) **distribution**, written in a symbol form, N(μ, 4).
- In many real-world applications, the known value, variance 4, is usually an educated guess taken from existing literature and poses as an assumption.
- Statement 2 is again taken as factual (because of the proof). It states that we can calculate the confidence interval from the right-hand side of equation 1.

- To conclude, unleash the power of confidence intervals! By mastering their art, you'll wield the ability to shape decisions, uncover truths, and make an impact.