Class 3b: Review of concepts in Probability and Statistics

Business Forecasting

Confidence Intervals

Confidence Intervals

- We calculated the mean price in our sample
- How confident are we that our estimate is close to the parameter's value?
- Confidence intervals measure uncertainty around the estimate

Confidence Intervals

- Mean price was 1245.43
- Is it reasonable to think true average price in population is 1100? What about 2000?
- Suppose that we calculated the confidence interval to be:

$$\{1086.64, 1404.22\}$$

  - Where are these numbers coming from?

1. The sampling distribution of the sample mean tells us how likely we are to get a point estimate which is far away from the true mean

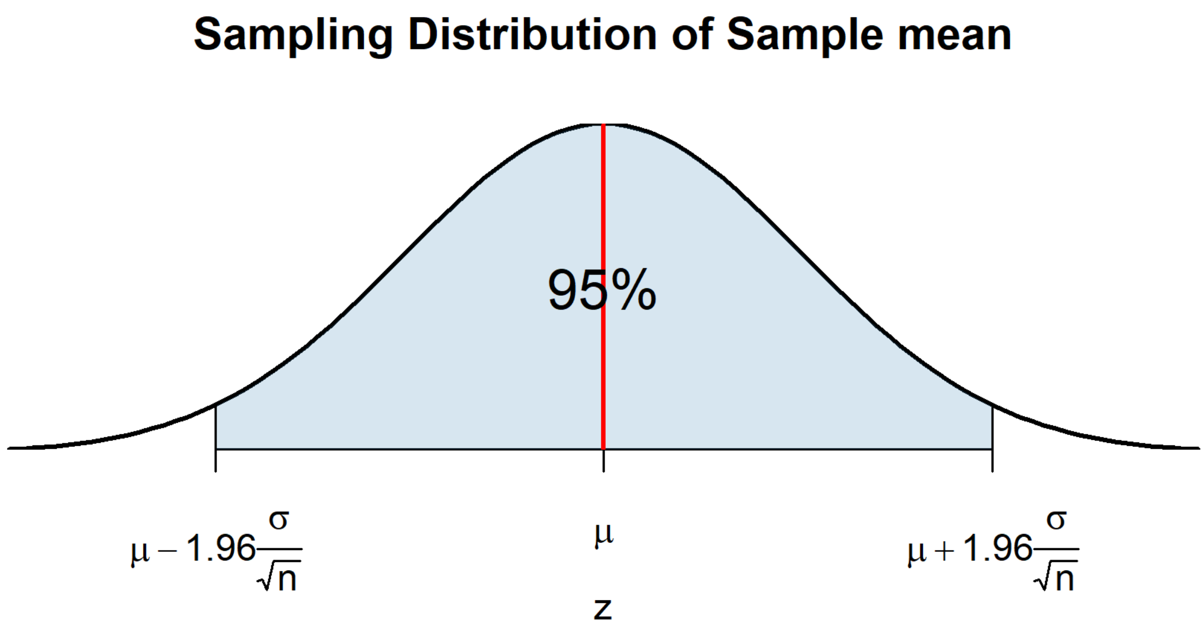
2. The confidence interval uses this property of the sampling distribution to tell us where the true mean might be

   - Let's go through these statements 1-by-1

Sampling distribution

Q: How likely is it that a sample mean is far away from the true mean?

- Consider a hypothetical sampling distribution of a sample mean
  - Reminder: $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$
  - If we draw samples repeatedly, 95% of their means will be within the shaded area
- Why 1.96?



Yet another way to see it

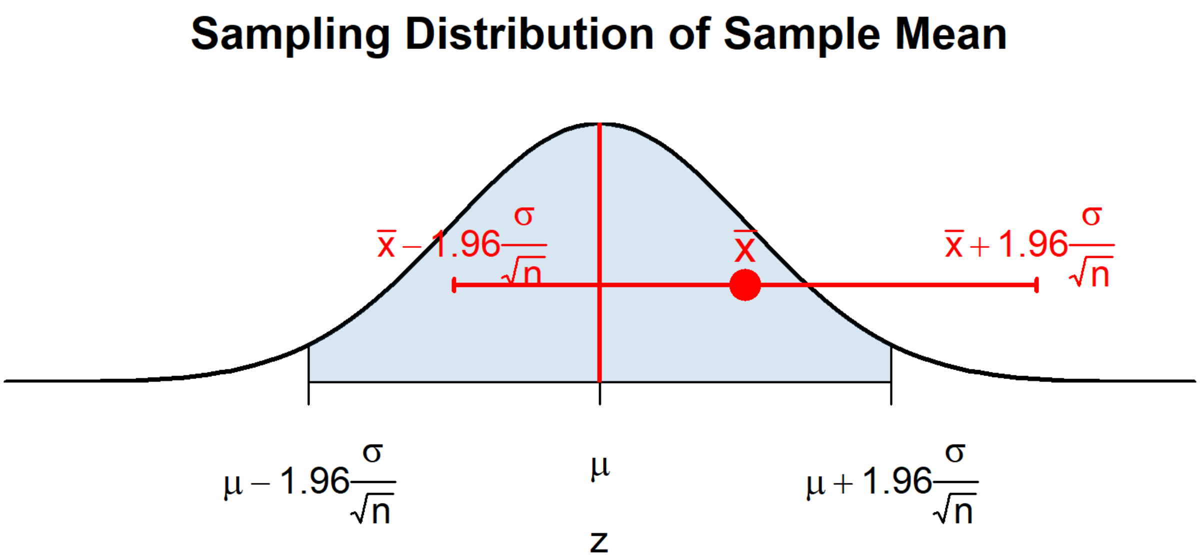
$$
\begin{aligned}
0.95 &= P(-1.96 < Z < 1.96) \\
&= P(z_{-\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) \\
&= P(z_{-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\frac{\alpha}{2}}) \\
&= P(z_{-\frac{\alpha}{2}}\sigma/\sqrt{n} < \bar{X} - \mu < z_{\frac{\alpha}{2}}\sigma/\sqrt{n}) \\
&= P(\mu - z_{-\frac{\alpha}{2}}\sigma/\sqrt{n} < \bar{X} < \mu + z_{\frac{\alpha}{2}}\sigma/\sqrt{n})
\end{aligned}
$$

- Theoretically, CLT theorem guarantees that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is standard normal
- What happens if you do not know σ?
- In large sample, $s \to \sigma$, so $\frac{\bar{X}-\mu}{s/\sqrt{n}} \to N(0,1)$
- So in large samples, standardized sample mean (with estimated standard deviation) will also have normal distribution
- You man need a bit higher n to ensure $s \to \sigma$

Sampling distribution

Q: How far is the sampled mean from the true mean?

- Hence 95% of the draws of sample means will be within distance of $1.96\frac{\sigma_X}{\sqrt{n}}$ to the true parameter
- There is only 5% chance that we have draw sample weird enough that $\bar{X}$ is further from μ_X by more than $1.96\frac{\sigma_X}{\sqrt{n}}$
- Confidence interval of $\bar{X}$ will cover μ_X as long as $|\mu_X - \bar{X}| < 1.96\frac{\sigma_X}{\sqrt{n}}$



Sampling distribution

- Suppose we draw many samples from the same distribution
- For each sample we compute the sample mean and we construct the interval
- 95% of them will cover the true population mean!

Source: [https://seeing-theory.brown.edu/frequentist-inference/index.html#section2)

Calculation Procedure

Use this procedure if $n > 40$

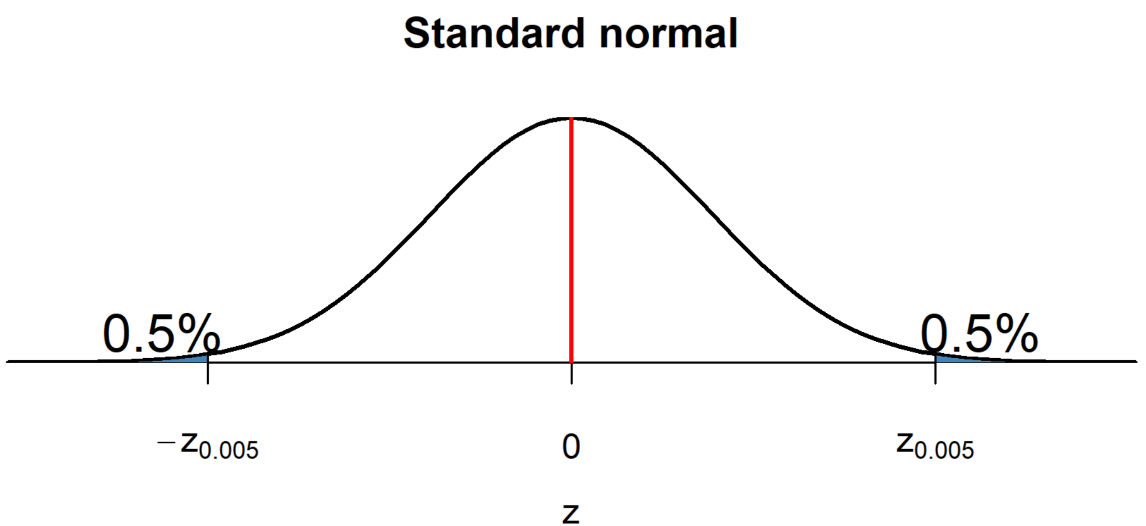
1. Take an IID sample
2. Calculate mean $\bar{x}$ and standard deviation s in your sample
   - Standard Error is standard deviation of the estimator $SE = \frac{s}{\sqrt{n}}$
3. Pick confidence level (usually 90,95,99%)
   - We typically denote the confidence level $1 - \alpha$
   - α is probability of making a Type 1 error (more about it later)
   - Example: if confidence level is 95%, $\alpha = 0.05$
4. Find the corresponding critical values $z_{\frac{\alpha}{2}}$
   - Critical values are such that $P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1 - \alpha$
   - Example: if confidence level is 95%, $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$
5. Construct the confidence interval as:

$$\{\bar{x} - z_{\frac{\alpha}{2}} * \underbrace{\frac{s}{\sqrt{n}}}_{SE}, \bar{x} + z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}\}$$

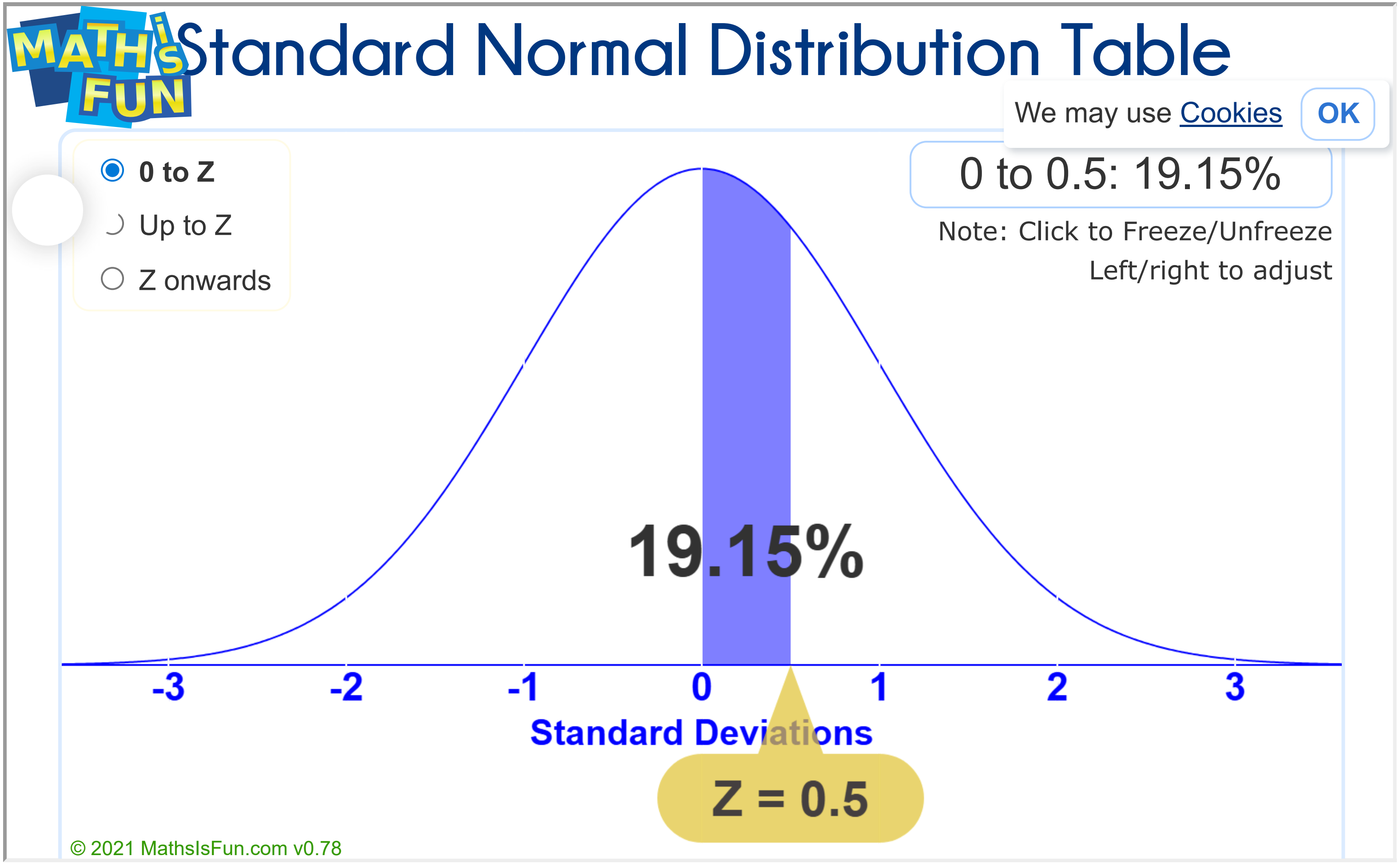
Finding Critical Values

- Suppose confidence interval is 99%.
- Then $\alpha = 0.01$
- We are looking for $z_{\frac{\alpha}{2}}$ such that:

$$P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 0.99$$



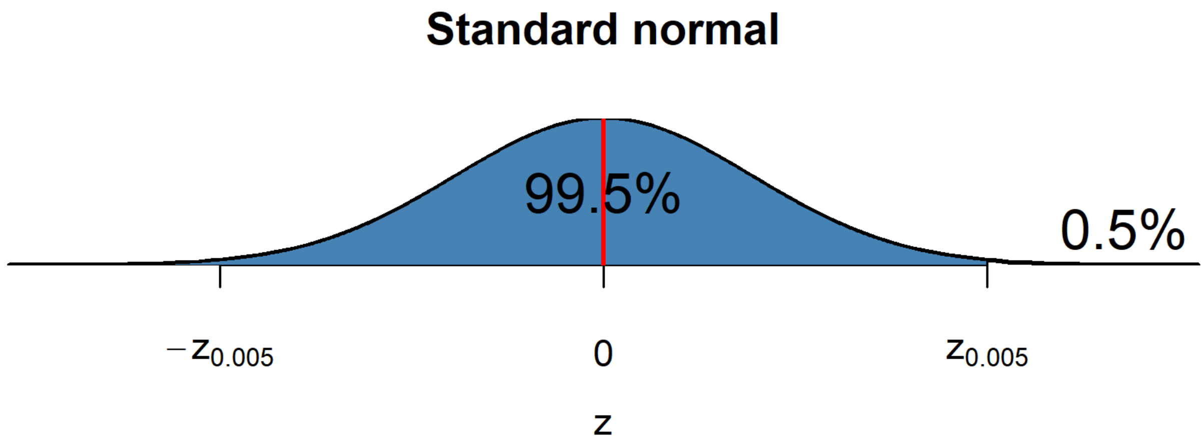
$$P(Z > z_{0.005}) = 0.005 \qquad \text{or} \qquad P(Z < z_{0.005}) = 0.995$$



Source: [https://www.mathsisfun.com/data/standard-normal-distribution-table.html)

Finding Critical Values

$P(Z < z_{\frac{\alpha}{2}}) = 0.995$ $z_{\frac{\alpha}{2}}$, is 99.5% quantile of standard normal $\rightarrow z_{\frac{\alpha}{2}} = 2.58$



Constructing CI: example

Let's calculate 90% CI for average price of listing with grade>4.5

1. Take an IID sample
   - $n = 100$ ✓
2. Calculate mean $\bar{x}$ and standard deviation s
   - $\bar{x} =$ 1245.43 and $s =$ 961.9

3. Pick confidence level

   - We pick 90%, so $\alpha = 0.1$

4. Find the corresponding critical values $z_{\frac{\alpha}{2}}$

   - Find $z_{\frac{\alpha}{2}}$ such that $P(Z > z_{\frac{\alpha}{2}}) = 0.05$ (or $P(Z < z_{\frac{\alpha}{2}}) = 0.95$)
   - $z_{0.05} = 1.65$
5. Construct the confidence interval as:

$$\{\bar{x} - z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}\}$$

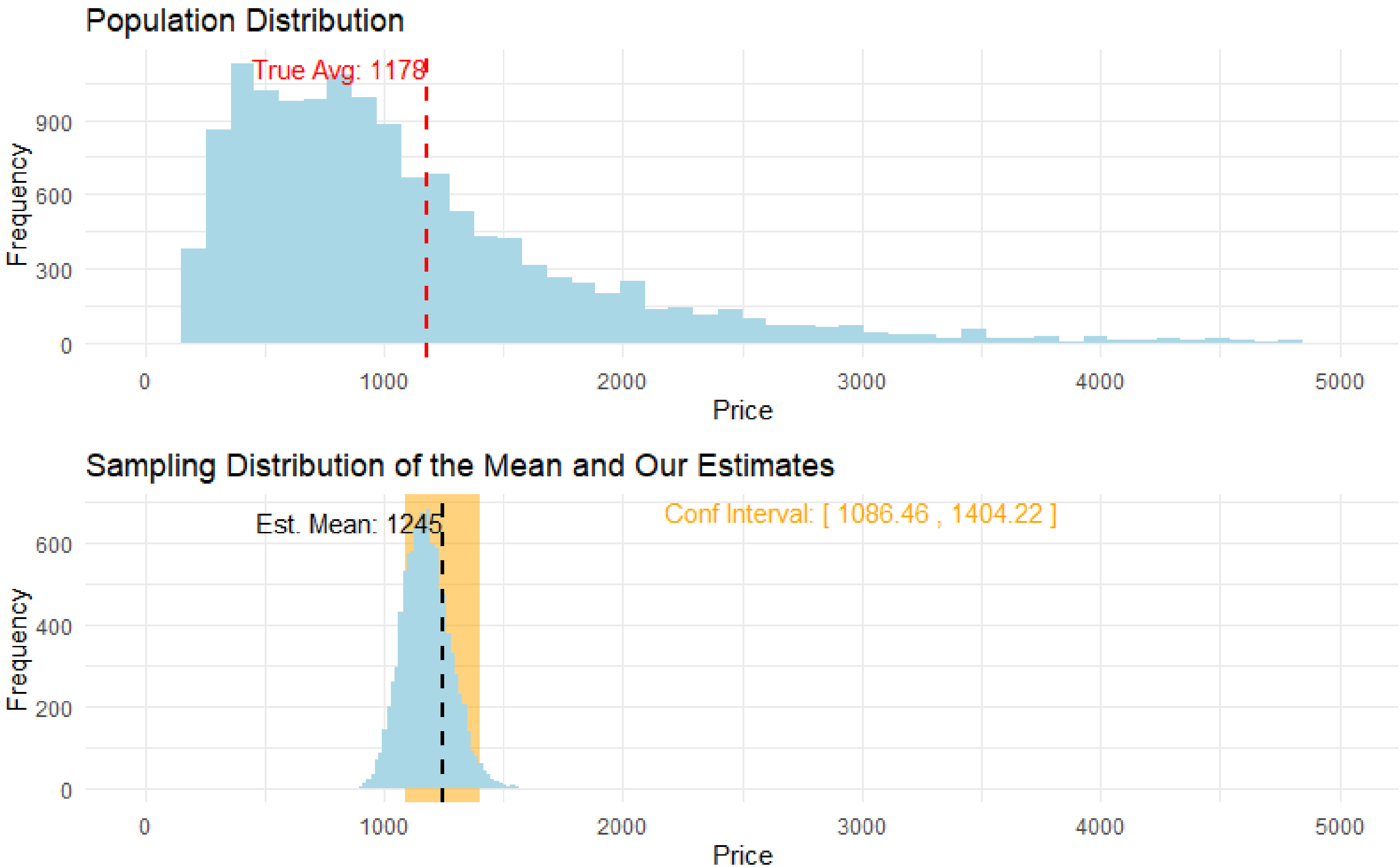
$$\{1245.43 - 1.65\frac{961.9}{\sqrt{100}}, 1245.43 + 1.65\frac{961.9}{\sqrt{100}}\}$$

Interpreting confidence intervals

$$CI_{90} = \{1086.64, 1404.22\}$$

How do we interpret a 90% confidence interval we computed?

- **Correct Interpretation**
  - We are 90% confident that the interval captures the true mean
  - We are 90% confident that the true mean price of listings with grade>4.5 is between 1086.64 and 1404.22
- **Incorrect**
  - With 90% probability the true mean is between 1086.64 and 1404.22
    - Computed interval is not-random and true mean is not random, so can't make probabilistic statements.
    - Interval is a function of random variables only **before** we draw a sample and make any computation.
    - After we have a sample, nothing is random. The true mean is either between 1086.64 and 1404.22 or not.



Population Distribution
True Avg: 1178
900
600
300
0
Frequency
0
1000
2000
3000
4000
5000
Price
Sampling Distribution of the Mean and Our Estimates
Est. Mean: 1245
Conf Interval: [1086.46 , 1404.22]
600
400
200
0
Frequency
0
1000
2000
3000
4000
5000
Price

Shape of confidence intervals

Confidence intervals $\{\bar{x} - z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}\}$ are wider when:

- Confidence level is higher (99% is wider than 90%)
- When n is small
- When σ is large

Practice

In learnr:

1. Set seed to to your student id.
2. Take a sample of 100 observations from the data
3. Calculate the mean and standard deviation of the sample
4. Find critical value
5. Compute confidence interval

What critical values?

When should we use critical values from Normal Distribution?

1. Original distribution (of X) is not normal:
   - If $n > 30$- use critical values from normal distribution (40 if σ unknown)
   - If $n < 30$ - you are screwed
2. Original distribution (of X) is normal:
   - If you know σ, you can use critical values from normal (n doesn't matter)
     - If X is normal, then use σ instead of s and $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
   - If you don't know σ but $n > 40$, you can use critical values from normal
     - CLT kicks in
   - If you don't know σ and $n < 40$, you use critical values from student's t.
     - $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ is not normal. s is not a good approx. of σ when n is low

What's Student's t?

If $X_1, X_2, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma)$, then

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

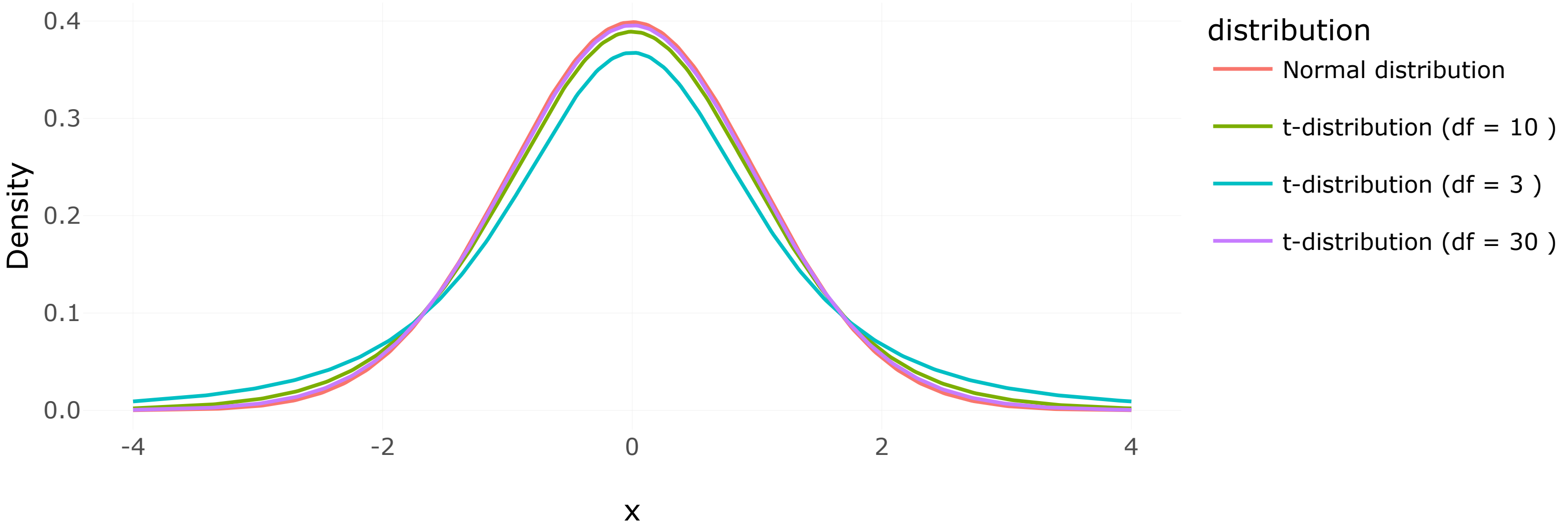
Where s is sample standard deviation.

T has a student's t distribution with n−1 degrees of freedom

$$T \sim t_{n-1}$$

What's Student's t?

- Bell shaped and symmetric around 0
- More spread out - heavier tails, more uncertainty (because we don't know standard deviation)
- Shape determined by the degrees of freedom.
    - As n increases (and hence degrees of freedom), it tends to standard normal (as it should by CLT!)
    - Less uncertainty because we are better at estimating standard deviation

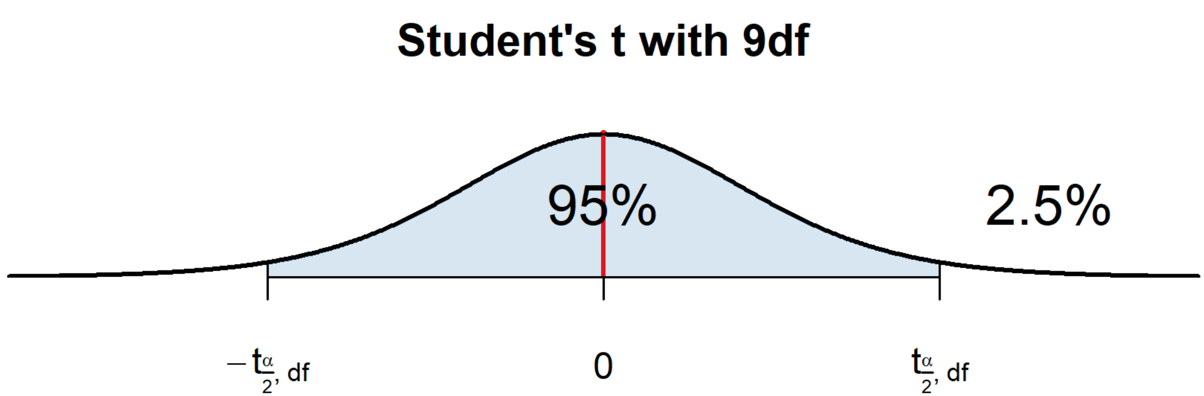


Student's t critical values

Finding critical values for student's t distribution:

1. Determine what is the right number of degrees of freedom ($n-1$)!
2. Determine what's your confidence level and your $(1-\alpha)$
   - From this figure out $\alpha/2$
3. Find the percentile such that

$$P(T > t_{\frac{\alpha}{2},\underbrace{n-1}_{d.f.}}) = \frac{\alpha}{2} \qquad \text{or} \qquad P(T < t_{\frac{\alpha}{2},n-1}) = 1 - \frac{\alpha}{2}$$

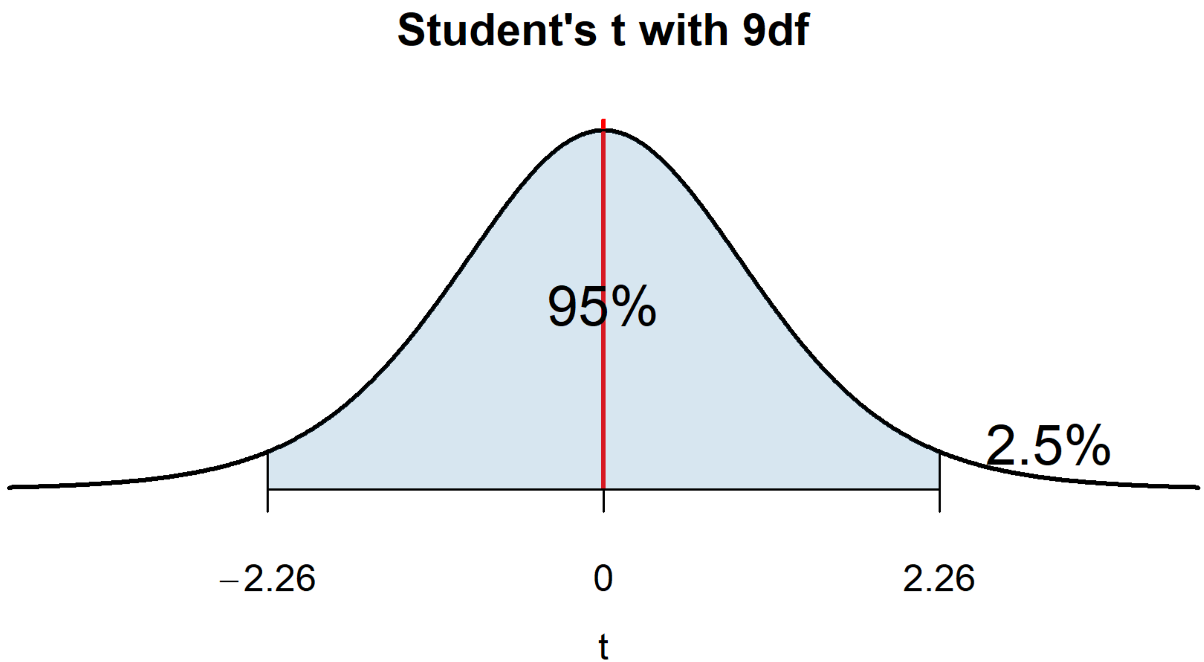


Example

- Practice in learnr finding critical values using the code:

```
qt(quantile, degrees_of_freedom)
#ex
qt(0.975, 9)
```

- $n = 10 \rightarrow df = 9$
- Confidence level is 95% $\rightarrow 1 - \alpha = 0.95$ and $\frac{\alpha}{2} = 0.025$
- What's $t_{0.025,9}$ such that $P(T < t_{0.025,9}) = 0.975$



- Once we have critical value, we construct the CI as before:

$$\{\bar{x} - t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}}\}$$

Practice in learnr:

Your company implemented free shipping for a random group of customers. They want to know whether it increased spending. Here is your data:

$157.80, $192.45, $210.20, $175.60, $198.30, $180.90, $205.75, $185.20, $177.40, $195.60

a) Calculate 90% confidence interval. What assumptions you need?

b) Average spending without free shipping is $182, can say anything about whether free shipping increased spending?

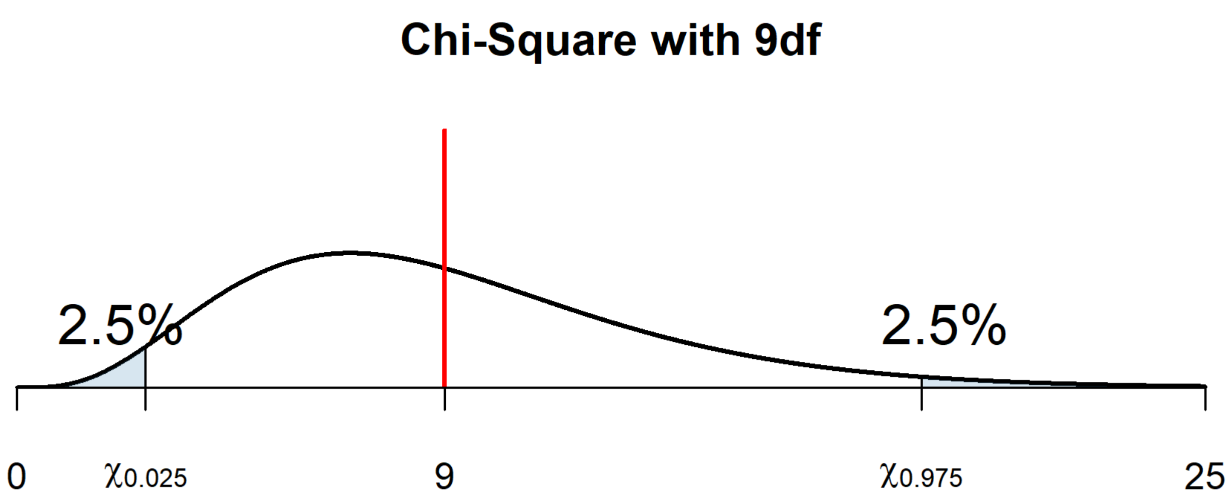
Confidence Intervals for Variance

- Suppose $X_1, X_2, \dots X_n$ come from normal distribution
- The sampling distribution of the sample variance $S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$ is:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}$$

- We will use the fact that:

$$P(\chi_{0.025,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi_{0.975,n-1}) = 0.95$$



Confidence Intervals for Variance

How we use it to construct the confidence interval?

$$
\begin{aligned}
0.95 &= P(\chi_{0.025,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi_{0.975,n-1}) \\
&= P(\frac{1}{\chi_{0.975,n-1}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi_{0.025,n-1}}) \\
&= P(\frac{(n-1)S^2}{\chi_{0.975,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi_{0.025,n-1}})
\end{aligned}
$$

So more generally, the confidence interval for the sample variance is

$$
CI_{1-\alpha} = \{\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2},n-1}}\}
$$

- Where $\chi_{1-\frac{\alpha}{2},n-1}$ and $\chi_{\frac{\alpha}{2},n-1}$ are quantiles of χ_{n-1} distribution, such that $P(X < \chi_{1-\frac{\alpha}{2},n-1}) = 1 - \frac{\alpha}{2}$ and $P(X < \chi_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2}$
- You can find them in R using:

```
qchisq(alpha, df)
```

Practice in learnr

Suppose you produce sausages. As a quality control, you measure the level of fat in your sausages. You take a random sample of 12 sausages and you find the variance of 20 ($grams^2$). Find 99% confidence interval for the variance. What assumptions you need?