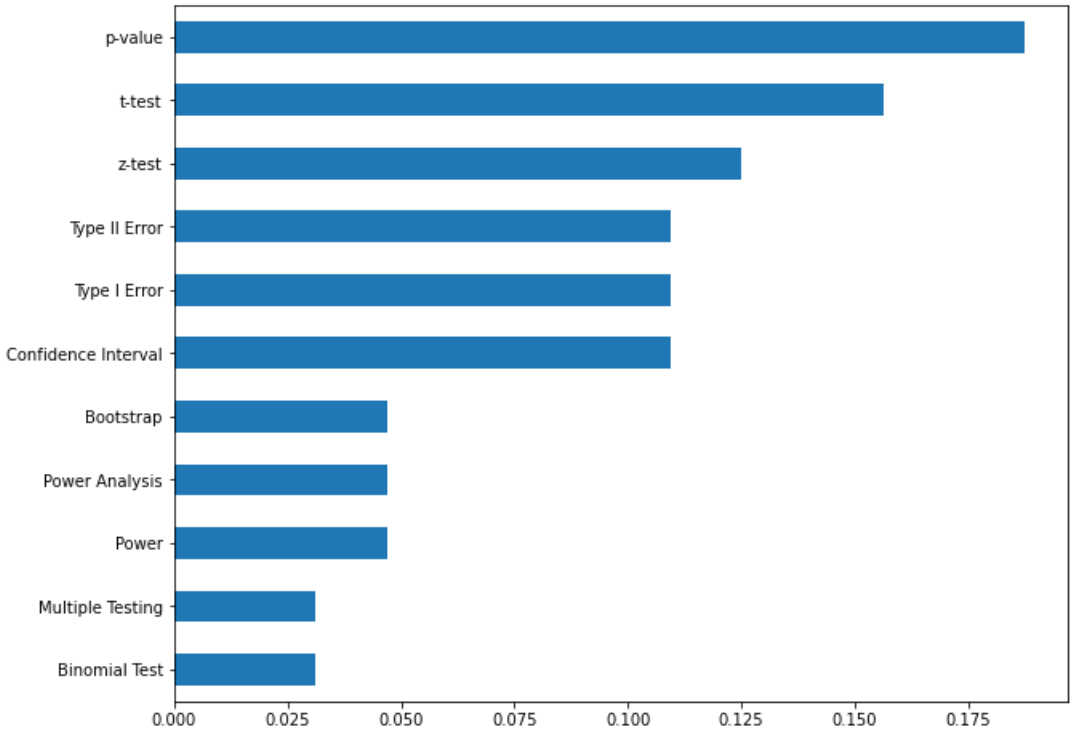
EMMA ding

emmading.com

info@emmading.com

One-sample t-tests

📍 T-test is one of the top topics asked in statistics interviews.



Lesson Structure

📍 Interview Questions

- What are differences between t-test and z-test?
- What are the assumptions of t-test?

▼ T-test vs. Z-test

- The z-test is used to test the **mean/proportion** of a population against a number or against the **mean/proportion** of another population, as long as some basic assumptions are satisfied.

  The data are **normally distributed with known variance**, or a large enough sample that we could invoke some well-known theorems (Central Limit Theorem, Slutsky's Theorem) to obtain a normally distributed test statistic.

- For a z-test or t-test with more than 30 sample points, no need to assume normality as the Central Limit Theorem implies that you can use a Z-test for both cases.
- For large sample sizes, the t-test procedure gives almost identical p-values as the Z-test procedure. The t-test can be replaced by a Z-test if we have over 30 samples (as long as the data is **not highly skewed**).
- Similarly to the Z-test, t-test also has one sample and two sample tests.

▼ Assumptions of t-tests

The t-test accomplishes the same goal as the z-test, but under a complementary set of assumptions:

- the **sample size is not too large** (usually less than 30)
- the **population variance is not known** (almost always the case in applications)
- the data is **normally distributed** (though we can relax this assumption slightly by requiring only that the sample mean be normally distributed, the sample variance be χ^2 distributed and independent of the sample mean).

> 📌 We should not use the t-test if the sample contains outliers as that may mean some of the required assumptions like normality are in fact violated.

▼ T-statistic

To use the t-test, the test statistic T

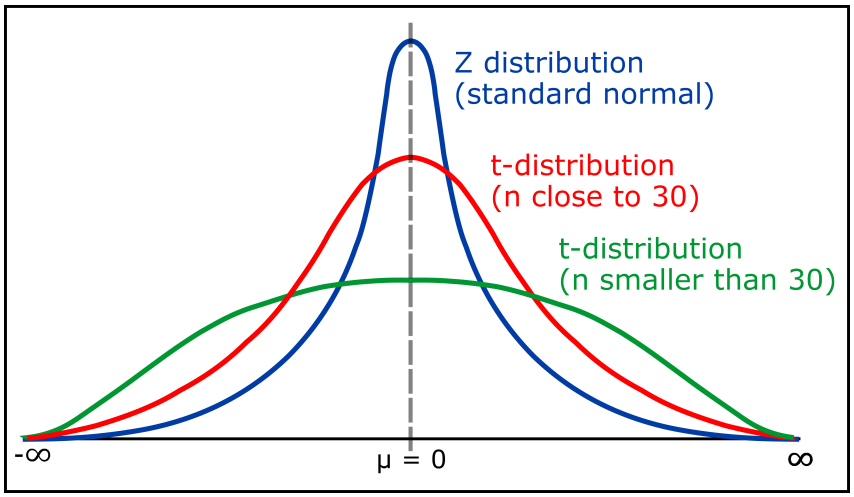
$$T = \frac{\text{sample mean} - \mu_0}{\text{sample standard deviation}}$$

Where μ_0 is a constant or the sample mean of a second population to which we would like to compare our population mean.

If the assumptions for the t-test are satisfied, then T follows the Student t-distribution.

▼ Student-t Distribution

T-distribution vs. z-distribution



Source https://www.geeksforgeeks.org/students-t-distribution-in-statistics/

▼ The dependence of the t-distribution on the sample size n is via its only parameter, which we call ***degrees of freedom***.

> ✏️ **Degrees of freedom**: The number of pieces of information that can freely vary without violating any given restrictions. That is, the number of independent pieces of information available to estimate another piece of information.
>
> - It's the number of **independent input** (e.g. the sample points) **minus** the number of **intermediate input** (e.g. the sample mean is viewed as a fixed parameter in the calculation of the sample variance).
> - The sample variance of n data points in the denominator of T depends on the sample mean, there're $n-1$ degrees of freedom.

- The shape of the t-distribution differs for sample sizes much smaller than 30 compared to sample sizes close to 30.

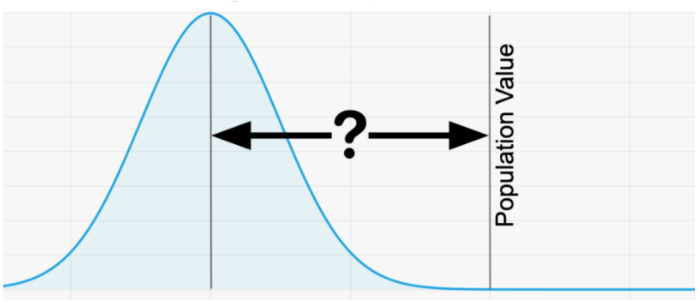
▼ The t-distribution has **heavier tails** than the normal distribution.

- Since we construct the test statistic T using the sample variance and not the population variance (which is unknown), we expect that T will take on extreme values more often that the Z-statistic.

- As n increases, the t-distribution better approximates the normal distribution. For large numbers of sample points – more than 30 - the t-distribution and the normal distribution become almost identical.

▼ T-test for One-Sample Mean

Let $X_1, ..., X_n$ be a small random sample $(n \leq 30)$ from a normal population with mean μ.



▼ **Hypothese**

$H_0 : \mu = \mu_0,$

$H_1 : \mu \neq \mu_0.$

▼ **Test statistic**

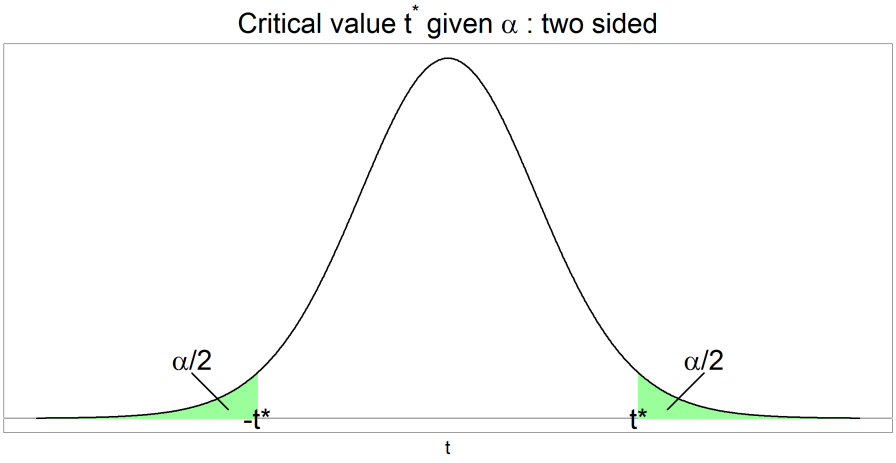
We can construct our t-statistic. Under H_0

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

Where $\bar{X}$ is the sample mean, μ_0 is a constant, and s is the sample standard deviation $s = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}}$. s^2 is an unbiased estimator of the variance.

▼ **Critical t value for a two-sided test**

Set significance level α, we define $t_{n-1,\alpha/2}$ the positive real number to the left of which an observed t_{n-1} random variable falls with probability $\Pr\left(T > t_{n-1,\alpha/2}\right) = \frac{\alpha}{2}$ and $\Pr\left(T < -t_{n-1,\alpha/2}\right) = \frac{\alpha}{2}$.



Note: $t^* = t_{n-1,\alpha/2}$

▼ **E.g. Estimate the average height of women.**

We want to estimate the average height of women in U.S. We randomly sampled 10 women. We guess the true value is around 160 cm.

Hypothese

$H_0 : \mu = 160,$

$H_1 : \mu \neq 160.$

```
import scipy.stats as stats

# simulate n height sample points
n = 10
sigma = 2.1
population_mean = 162
x = stats.norm.rvs(loc=population_mean, scale=sigma, size=n, random_state=1)

# Test if mean height is close to mu_0 using a T-test with 95% CL
mu_0 = 160
observed_t_score = (x.mean() - mu_0) / (x.std() / n**0.5)
critical_t_score = stats.t.ppf(0.975, n-1)
print('observed_t_score = ', observed_t_score)
print('critical_t_score = ', critical_t_score)
```

```
observed_t_score =  2.2709797875831743
critical_t_score =  2.2621571627409915
```

We land in the rejection region, so we reject our null hypothesis in favor of the alternative: the mean height is likely different from 160 cm.

▼ **Small-sample confidence interval for a population mean**

A $(1 - \alpha) * 100\%$ confidence interval for μ is:

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

If the sample size is large, we could instead use: $\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

```
# Build confidence interval at 5% significance level for the population mean
margin_of_error = critical_t_score * (x.std() / n**0.5)
lower = x.mean() - margin_of_error
upper = x.mean() + margin_of_error
confidence_interval = (lower, upper)
print(confidence_interval)
```

```
(160.00697737194074, 163.58503088667365)
```