# Fundamental Sampling Distributions

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy.stats import norm, poisson, chi2
from scipy.stats import t as student_t
from scipy.stats import f as f_dist
from scipy.signal import fftconvolve
```

## Recap on Populations and Samples

We begin this chapter by discussing the notions of *populations* and *samples*. Both are mentioned in a broad fashion in Chapter 0.1 and Chapter 0.2. However, much more needs to be presented about them here.

> A **population** consists of the totality of the observations with which we are concerned.

The number of observations in the population is called the **size of the population**. A population may be of finite size or infinite size. For examples,

- the numbers on the cards in a deck,
- the heights of residents in a city at certain date and time,
- the weights of fish in a particular lake at certain date and time

are all finite sized populations. On the other hand, the observations such as

- atmospheric pressure every day, from the past on into the future,
- all measurements of the depth of a lake, from any conceivable position

are examples of infinite sized populations.

In practical, some finite populations are so large that we assume them to be infinite. This is true in the case such as the population of lifetimes of a certain type of storage battery being manufactured for mass distribution throughout the country.

**Each observation in a population is a value of a random variable** $X$ having some probability distribution $f(x)$. For examplem If one is inspecting items coming off an assembly line for defects, then each observation in the population might be a value $0$ or $1$ of the

Bernoulli random variable $X$ with probability distribution $b(x;\ 1, p)$ where $0$ indicates a non-defective item and $1$ indicates a defective item.

When we refer hereafter to a "*binomial population*," a "*normal population*," or, in general, the "*population $f(x)$*," we shall mean a population whose observations are values of a random variable having a binomial distribution, a normal distribution, or the probability distribution $f(x)$.

In the field of statistical inference, we are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. For example, in attempting to determine the average length of life of a certain brand of light bulb, it would be impossible to test all such bulbs if we are to have any left to sell. Therefore, we must depend on a smaller subset of observations from the population to help us make inferences concerning that same population. This brings us to consider the notion of *sampling*.

> A **sample** is a subset of a population.

If our inferences from the sample to the population are to be valid, we must obtain samples that are **representative** of the population. We already discussed about these things in detail in Chapter 0.2. To recap, to eliminate any possibility of **bias** in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.

> The rv's $X_1, X_2, \ldots, X_n$ are said to form a **random sample** of size $n$ if
>
> 1. Every $X_i$ has the same probability distribution,
>
> $$X_i \sim f(x) \quad \text{for all } x = 1, 2, \ldots, n.$$
>
> 2. The $X_i$'s are independent rv's, so the joint probability of them can be written as
>
> $$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$
>
> Or we can say that the $X_i$'s are *independent and identically distributed* (iid).

## Some Important Statistics

Our main purpose in selecting random samples is to elicit information about the unknown **population parameters**.

Suppose, for example, that we wish to arrive at a conclusion concerning the proportion of coffee-drinkers in the US who prefer a certain brand of coffee. It would be impossible to question every coffee-drinking American in order to compute the value of the parameter $p$ representing the proportion. Instead, a large random sample is selected and the proportion $\hat{p}$ of people in this sample favoring the brand of coffee in question is calculated. The value $\hat{p}$ is now used as an *estimate* to make an inference to the true proportion $p$.

Now, **$\hat{p}$ is a function of the the random sample**; since many random samples are possible from the same population, we would expect $\hat{p}$ to vary somewhat from sample to sample. That is, **$\hat{p}$ is a value of a random variable** that we represent by $P$. Such a random variable is called a *statistic*.

> Any function of the random variables constituting a random sample is called a **statistic**.

We already introduced some important statistics (we called it numerical summary measures back then) in Chapter 0.1. Here, we will quickly recap them and introduce some notations for them in the context of random variable.

---

## Location Measures of a Sample

Let $X_1, X_2, \ldots, X_n$ be random variables representing a random sample of size $n$:

> (a) The **sample mean** is
>
> $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

> (b) The **sample median** is
>
> $$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{x_{n/2+1}}) & \text{if } n \text{ is even.} \end{cases}$$

> (c) The **sample mode** is the value of the sample that occurs most often.

---

## Variability Measures of a Sample

Agian, let $X_1, X_2, \ldots, X_n$ be random variables representing a random sample of size $n$:

(a) The **sample variance** is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

For alternative formulation, we can write

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}X_i^2 - \left(\sum_{i=1}^{n}X_i\right)^2\right].$$

(b) The **sample standard deviation** is

$$S = \sqrt{S^2},$$

where $S^2$ is the sample variance.

(c) The **sample range** is

$$R = X_{\max} - X_{\min}.$$

## Sampling Distributions

Since a statistic is a random variable that depends only on the observed sample, it must have a probability distribution.

The probability distribution of a statistic is called a **sampling distribution**.

The sampling distribution of a statistic depends on the distribution of the population, the size of the samples, and the method of choosing the samples. For instance, The probability distribution of $\bar{X}$ is called the **sampling distribution of the mean**.

## Distribution of Sample Means & The Central Limit Theorem

The first important sampling distribution to be considered is that of the mean $\bar{X}$. Suppose that a random sample of $n$ observations is taken from a population of probability distribution $f(x)$ with mean $\mu$ and variance $\sigma^2$. Then the sample mean of this sample, given by

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n),$$

so the **mean of the sample mean** is given by

$$\begin{aligned}\bar{x} = \mu_{\bar{X}} = E(\bar{X}) &= E\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\ &= \frac{1}{n}[E(X_1) + E_{(}X_2) + E(X_n)] \\ &= \frac{1}{n}(\underbrace{\mu + \mu + \cdots + \mu}_{n \text{ terms}}) = \mu\end{aligned}$$

and the **variance of the sample mean** is

$$\begin{aligned}s^2 = \sigma^2_{\bar{X}} &= \mathrm{Var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) \\ &= \frac{1}{n^2}[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + cdots + \mathrm{Var}(X_n)] \\ &= \frac{1}{n^2}(\underbrace{\sigma^2 + \sigma^2 + \cdots + \sigma^2}_{n \text{ terms}}) = \frac{\sigma^2}{n}.\end{aligned}$$

Thus, the mean of sample mean $\mu_{\bar{X}}$ is (theoretically) equal to the population mean $\mu$. We can also see the variance of sample mean $\sigma_{\bar{X}} = \sigma^2/n$ is smaller when $n$ is getting larger, implying that a larger sample size makes variability of the sample mean smaller and in turn makes the sample mean a better estimation to the population mean.

In addition, if we consider the **sample total** $T_n$ given by

$$T_n = X_1 + X_2 + \cdots + X_n = n\bar{X}$$

then the mean and variance of the sample total is $\mu = n\mu$ and $\sigma^2 = n\sigma^2$ respectively.

---

## The Case of Normal Populations

> If $X_1, X_2, \ldots, X_n$ is a random sample from a **normal population** with a normal distribution with mean $\mu$ and standard deviation $\sigma$. Then for any $n$, the sample mean $\bar{X}$ **is normally distributed** with mean $\mu$ and stadard deviation $\sigma/\sqrt{n}$.
>
> $$\bar{X} \sim N(\mu, \sigma/\sqrt{n}) \quad \text{if} \quad X_i \sim N(\mu, \sigma)$$

The proposition above also applied to the sample total $T_n$ but with mean $n\mu$ and standard deviation $\sqrt{n}\sigma$. This result is called the **reproductive property** of normal distribution. The term reproductive refers to the property of a distribution that when adding two or more independent rv's of that same particular distribution, the result will have the same distribution as the original rv's. Do note that some other distributions such as binomial, Poisson, and chi-squared also have this property.

**Example** An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Since the bulbs is a sample from approximately normal population, the sample mean $\bar{X}$ is normally distributed with mean $\mu_X = 800$ hours and standard deviation $\sigma_{\bar{X}} = 40/\sqrt{16} = 10$ hours. The desire probability is the area under $N(\bar{x};\ 800, 16)$ to the left of 775 as given by

$$P(\bar{X} < 775) = \int_{-\infty}^{775} N(\bar{x};\ 800, 10)\, d\bar{x}$$

The z-score corresponding to $\bar{x} = 775$ is

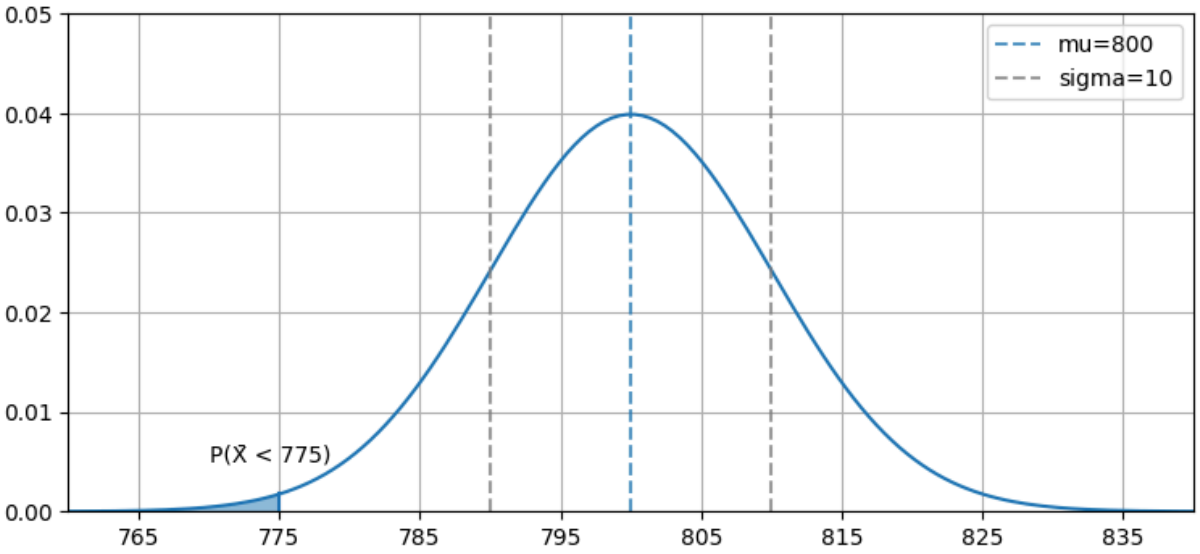
$$z = \frac{775 - 800}{10} = -2.5,$$

and therefore

$$P(\bar{X} < 775) = P(Z \leq -2.5) = 0.00621.$$

In [2]:
```
fig, ax = plt.subplots(1,1,figsize=(9,4))

xs = np.linspace(760,840,150)
xs_ = np.linspace(760,775,100)

ax.plot(
    xs, norm.pdf(xs, 800, 10)
)
ax.fill_between(
    xs_, xs_ * 0, norm.pdf(xs_, 800, 10),
    alpha=0.5
)
ax.plot([775,775], [0,norm.pdf(775, 800, 10)], color='tab:blue')
ax.plot([800,800], [0,0.05], color='tab:blue', linestyle='--', alpha=0.75, label='m
ax.plot([810,810], [0,0.05], color='tab:grey', linestyle='--', alpha=0.75, label='s
ax.plot([790,790], [0,0.05], color='tab:grey', linestyle='--', alpha=0.75)
ax.text(770, 0.005, "P(X̄ < 775)")

ax.set(
    xlim=[760,840], ylim=[0,0.05], axisbelow=True,
    xticks=np.arange(765,836,10)
```

```
)
ax.grid(); ax.legend();
```




---

## The Central Limit Theorem

If we are sampling from a population with unknown (non-normal) distribution, either finite or infinite, the sampling distribution of $\bar{X}$ will still be *approximately normal* with mean $\mu$ and variance $\sigma^2/n$, provided that the sample size is large. This fascinating result is an immediate consequence of the following theorem, called the *Central Limit Theorem*.

> **The Central Limit Theorem (CLT)**
>
> Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from a distribution with mean $\mu$ and variance $\sigma^2$. Then, if $n$ is sufficiently large, the sample mean $\bar{X}$ is approximately normally distributed with mean $\mu_{\bar{X}} = \mu$ and variance $\sigma^2_{\bar{X}} = \sigma^2/n$.
>
> In the limiting form, we can say that the probability distribution of
>
> $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$
>
> as $n \to \infty$ is the standard normal distribution $N(z;\ 0, 1)$.

# The proof of this theorem is sketched at the end of this section.

As a rule of thumb, the normal approximation for $\bar{X}$ will generally be good if $n \geq 30$, provided the population distribution is not terribly skewed. If $n < 30$, the approximation is good only if the population is not too different from a normal distribution and, as stated in previous section, if the population is known to be normal, the sampling distribution of $\bar{X}$ will follow a

normal distribution exactly, no matter how small the size of the samples. The same principle also apply for sample total $T_n$.

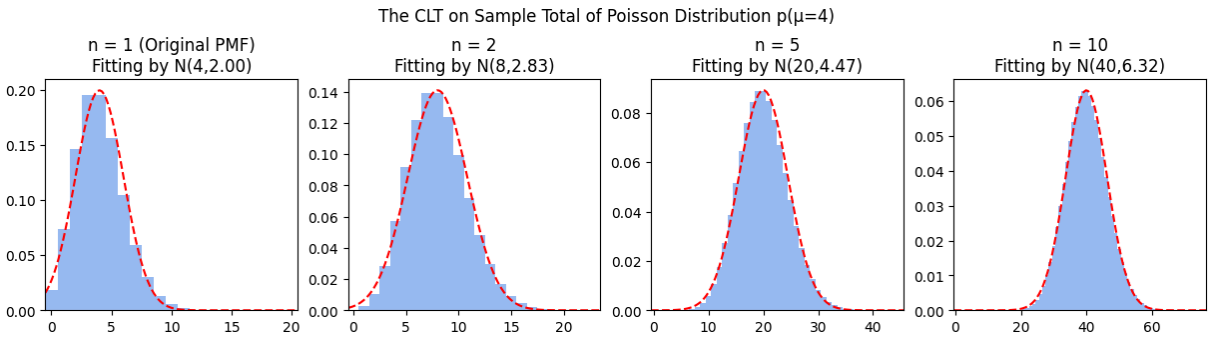
One application of the CLT is that it can be used to justify the normal approximation to binomial distribution, since a binomial random variable $X \sim b(n,p)$ is a result from $n$ successive independent trials and each trial can be though of as a Bernoulli random variable $X_i \sim b(1,p),\ i = 1, 2, \ldots, n$, so we can write

$$X = X_1 + X_2 + \cdots + X_n,$$

thus, a binomial rv is simply a sample total of Bernoulli population and so by the CLT, for large value of $n$, the binomial distribution is approximately a normal distribution.

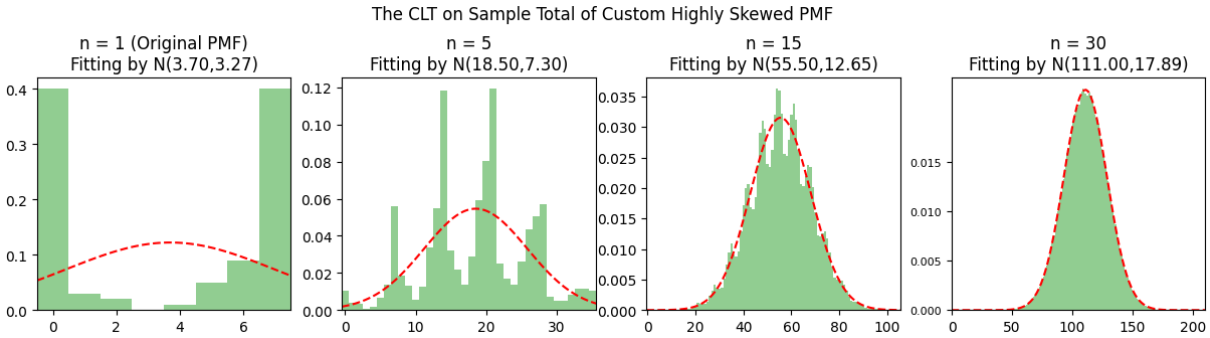
The following figures illustrate the Central Limit Theorem by showing the sampling distribution of sample means from various distribution and various sample size $n$, we can see that all sampling distribution turn into nearly bell-curve for large value of $n$. (Do note that the figures show the distribution of sample total instead of sample mean to simplify the calculation needed to create them.)

In [3]:
```
fig, axs = plt.subplots(1,4,figsize=(15,3))

fig.suptitle("The CLT on Sample Total of Poisson Distribution p(μ=4)", y=1.12)

Y = [poisson.pmf(np.arange(0,21), 4)]
for i in range(2,11):
    Y_ = np.array([fftconvolve(Y[i-2], Y[0])])
    Y = Y + list(Y_[:, :int(4*i + 6*(2*np.sqrt(i)))])

ns = [1,2,5,10]
for i,n in enumerate(ns):
    xs = np.arange(0, Y[n-1].size)
    axs[i].bar(xs, Y[n-1], width=1, alpha=0.65, color='cornflowerblue')
    xs = np.linspace(-0.5, Y[n-1].size, 120)
    axs[i].plot(xs, norm.pdf(xs, loc=4*n, scale=2*np.sqrt(n)), 'r--')
    axs[i].set(
        title = f"n = {n}" + (" (Original PMF)" if n == 1 else "") + f"\nFitting by
        xlim = [-0.5, Y[n-1].size-0.5]
    )
```




In [4]:
```
fig, axs = plt.subplots(1,4,figsize=(15,3))
```

```
fig.suptitle("The CLT on Sample Total of Custom Highly Skewed PMF", y=1.12)

Y = [np.array([0.4, 0.03, 0.02, 0, 0.01 ,0.05, 0.09, 0.4])]
mu = sum([p*i for i,p in enumerate(Y[0])])
sigma = np.sqrt(sum([(i - mu)**2 * p for i,p in enumerate(Y[0])]))
for i in range(2,51):
    Y_ = np.array([fftconvolve(Y[i-2], Y[0])])
    Y = Y + list(Y_[:, :int(mu*i +6*(sigma*np.sqrt(i)))])

ns = [1,5,15,30]
for i,n in enumerate(ns):
    xs = np.arange(0, Y[n-1].size)
    axs[i].bar(xs, Y[n-1], width=1, alpha=0.5, color='tab:green')
    xs = np.linspace(-0.5, Y[n-1].size, 120)
    axs[i].plot(xs, norm.pdf(xs, loc=mu*n, scale=sigma*np.sqrt(n)), 'r--')
    axs[i].set(
        title = f"n = {n}" + (" (Original PMF)" if n == 1 else "") + f"\nFitting by
        xlim = [-0.5, Y[n-1].size-0.5]
    )

axs[3].set_yticks([0, 0.005, 0.01, 0.015])
axs[3].tick_params(axis='y', labelsize=8)
```




---

## Examples of Inferences on the Population Mean

**Example** **Automobile Parts**: An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produce parts having a mean diameter of $5.0$ millimeters. The engineer involved conjectures that the population mean is $5.0$ millimeters. An experiment is conducted in which $100$ parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation is $\sigma = 0.1$ millimeter. The experiment indicates a sample average diameter of $\bar{x} = 5.027$ millimeters. Does this sample information appear to support or refute the engineer's conjecture?

Whether the data support or refute the conjecture depends on the probability that data similar to those observed in this experiment ($\bar{x} = 5.027$) can readily occur when in fact $\mu = 5.0$. In other words, we want to find how likely is it that one can obtain $\bar{x} \geq 5.027$ with $n = 100$ if the population mean is $\mu = 5.0$. Namely we want to find

$$P(|\bar{X} - 5| \geq 0.027),$$

which is the chance that $\bar{X}$ will deviate as much as $0.027$ millimeter (which is the deviation that we observed in the experiment). Continuing on the calculation:

$$P(|\bar{X} - 5| \geq 0.027) = P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq 0.027)$$

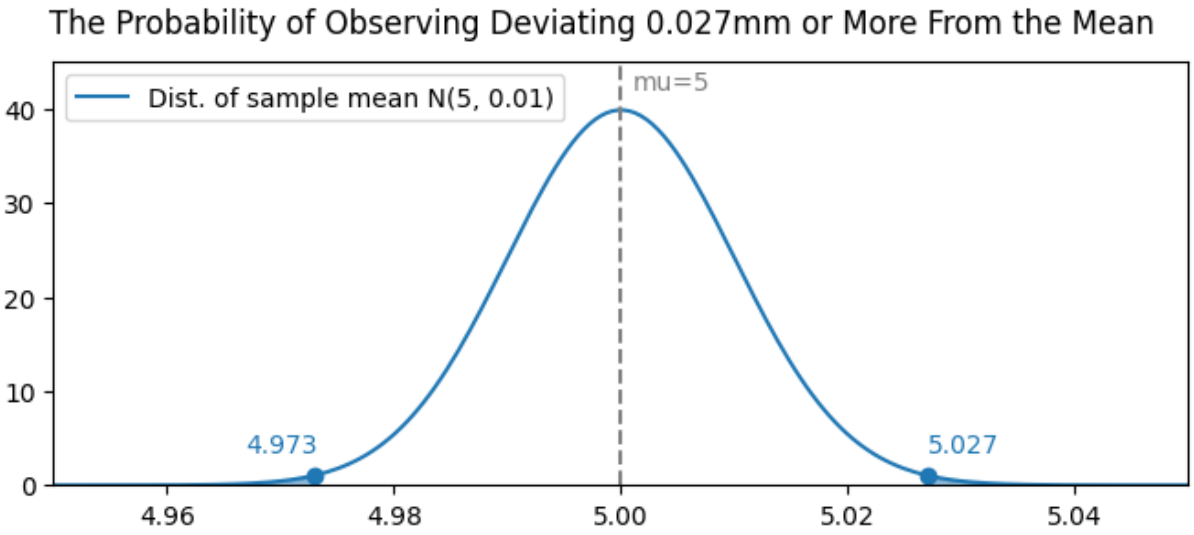
By the CLT, we assume that $\bar{X}$ is normally distributed, so the probability is symmetric around the mean:

$$\begin{aligned} P(|\bar{X} - 5| \geq 0.027) &= P(\bar{X} - 5 \geq 0.027) + P(\bar{X} - 5 \leq 0.027) \\ &= 2P(\bar{X} - 5 \geq 0.027) \\ &= 2P\left(\frac{\bar{X} - 5}{0.1/\sqrt{100}} \geq \frac{0.027}{0.1/\sqrt{100}}\right) \\ &= 2P(Z \geq 2.7) \approx 0.007 \end{aligned}$$

Therefore, one would experience by chance that an $\bar{x}$ would be 0.027 millimeter from the mean in only around 7 in 1000 experiments (0.7%). As a result, this experiment with $\bar{x} = 5.027$ certainly does not give supporting evidence to the conjecture that $\mu = 5.0$. In fact, it strongly refutes the conjecture!

In [5]:
```
fig, ax = plt.subplots(1,1,figsize=(8,3))

fig.suptitle("The Probability of Observing Deviating 0.027mm or More From the Mean"

xs = np.linspace(4.95, 5.05, 200)
ys = norm.pdf(xs, 5, 0.01)

ax.plot(
    xs, ys, label="Dist. of sample mean N(5, 0.01)"
)
ax.plot([5,5], [0,45], color='grey', linestyle='--')

xs_ = np.linspace(4.95, 5 - 0.027, 200)
ys_ = norm.pdf(xs_, 5, 0.01)
ax.fill_between(
    xs_, xs_*0, ys_,
    color='tab:blue', alpha=0.5
)
ax.plot([5 - 0.027, 5 - 0.027], [0, norm.pdf(5 - 0.027, 5, 0.01)], color='tab:blue'

xs_ = np.linspace(5 + 0.027, 5.05, 200)
ys_ = norm.pdf(xs_, 5, 0.01)
ax.fill_between(
    xs_, xs_*0, ys_,
    color='tab:blue', alpha=0.5
)
ax.plot([5 + 0.027, 5 + 0.027], [0, norm.pdf(5 + 0.027, 5, 0.01)], color='tab:blue'

ax.scatter([5 - 0.027, 5 + 0.027], norm.pdf([5 - 0.027, 5 + 0.027], 5, 0.01))
```

```
ax.text(5.001, 42, "mu=5", color='grey')
ax.text(4.975 - 0.008, 3.5, "4.973", color='tab:blue')
ax.text(5.027, 3.5, "5.027", color='tab:blue')

ax.set(
    xlim=[4.95, 5.05], ylim=[0, 45]
)
ax.legend();
```




**Example** Traveling between two campuses of a university in a city via shuttle bus takes, on average, $28$ minutes with a standard deviation of $5$ minutes. In a given week, a bus transported passengers $40$ times. What is the probability that the average transport time was more than $30$ minutes? Assume the mean time is measured to the nearest minute.

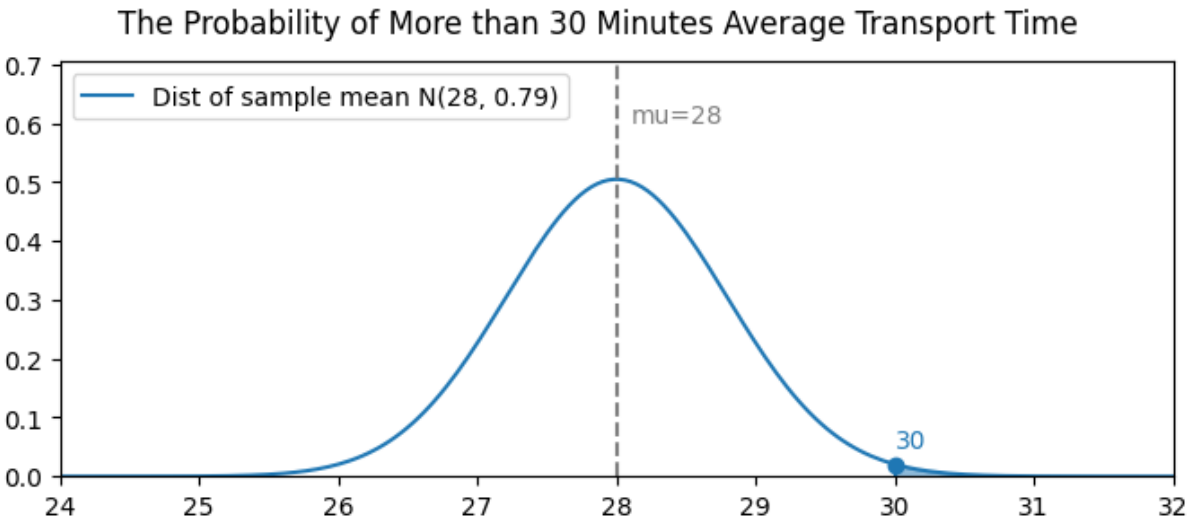
In this case, the population parameters are $\mu = 28$ and $\sigma = 5$, we want to find the probability that $\bar{X} > 30$ given that $n = 40$. Since the time is measured on a continuous scale to the nearest minute, an $\bar{X}$ greater than $30$ is equivalent to $\bar{x} \geq 30.5$. Hence,

$$P(\bar{X} \geq 30) = P\left(\frac{\bar{X} - 28}{5/\sqrt{40}} \geq \frac{30.5 - 28}{5/\sqrt{40}}\right) = P(Z \geq 3.16) = 0.0008$$

So there is only 0.08% chance that that the average time of one bus trip will exceed $30$ minutes.

In [6]:
```
fig, ax = plt.subplots(1,1,figsize=(8,3))

fig.suptitle("The Probability of More than 30 Minutes Average Transport Time")

xs = np.linspace(24, 32, 200)
ys = norm.pdf(xs, 28, 5/np.sqrt(40))

ax.plot(
    xs, ys, label="Dist of sample mean N(28, 0.79)"
)
ax.plot([28,28], [0,1], color='grey', linestyle='--')
```

```
xs_ = np.linspace(30, 32, 200)
ys_ = norm.pdf(xs_, 28, 5/np.sqrt(40))
ax.fill_between(
    xs_, xs_*0, ys_,
    color='tab:blue', alpha=0.5
)
ax.plot([30, 30], [0, norm.pdf(30, 28, 5/np.sqrt(40))], color='tab:blue', linestyle

ax.scatter([30], norm.pdf(30, 28, 5/np.sqrt(40)))

ax.text(28.1, 0.6, "mu=28", color='grey')
ax.text(30, 0.05, "30", color='tab:blue')

ax.set(
    xlim=[24, 32], ylim=[0, ys.max() + 0.2]
)
ax.legend();
```




---

## Difference Between Two Means

The Central Limit Theorem can be extended to the two-sample, two-population case.

> If independent samples of size $n_1$ and $n_2$ are drawn at random from two populations, discrete or continuous, with means $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively, then the sampling distribution of the **differences of means**, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed for large values of $n_1$ and $n_2$ with mean and variance given by
>
> $$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 \quad \text{and} \quad \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

> Hence,
>
> $$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$
>
> is approximately a standard normal variable.

**Example** The television picture tubes of manufacturer $A$ have a mean lifetime of $6.5$ years and a standard deviation of $0.9$ year, while those of manufacturer $B$ have a mean lifetime of $6.0$ years and a standard deviation of $0.8$ year. What is the probability that a random sample of $36$ tubes from manufacturer $A$ will have a mean lifetime that is at least $1$ year more than the mean lifetime of a sample of $49$ tubes from manufacturer $B$?

We are given the following information:

| Population $A$ | Population $B$ |
|---|---|
| $\mu_A = 6.5$ | $\mu_B = 6.0$ |
| $\sigma_A = 0.9$ | $\sigma_B = 0.8$ |
| $n_A = 36$ | $n_B = 49$ |

From the theorem, the sampling distribution of $\bar{X}_A - \bar{X}_B$ is approximately normally distributed with mean and standard deviation

$$\mu_{\bar{X}_A - \bar{X}_B} = 6.5 - 6.0 = 0.5 \quad \text{and} \quad \sigma_{\bar{X}_A - \bar{X}_B} = \sqrt{\frac{0.9^2}{36} + \frac{0.8^2}{49}} = 0.189$$

We interested in the probability $P(\bar{X}_A - \bar{X}_B \geq 1)$, to calculate it, consider the z-score corresponding to that value:

$$z = \frac{1.0 - 0.5}{0.189} = 2.65$$

Therefore,

$$\begin{aligned} P(\bar{X}_A - \bar{X}_B \geq 1) &= P(Z \geq 2.65) \\ &= 1 - P(Z < 2.65) = 0.004. \end{aligned}$$

---

## Sample Proportion

Consider a Bernoulli population where each observation can be classify into two type, generally labeled as "success" and "failure." We can defined a *sample proportion* $\hat{p}$ to be a

statistics representing the proportion of "success" observations in a random sample. More formally, we can defined $\hat{p}$ as follows:

> Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ drawn from a *Bernoulli population* such that
>
> $$X_i = \begin{cases} 1, & \text{if the } i\text{-th individual has the characteristic of interest} \\ 0, & \text{otherwise} \end{cases}$$
>
> The **sample proportion $\hat{p}$** is defined as the sample mean of these Bernoulli sample:
>
> $$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

This statistic estimates the population proportion $p$, which is the probability that a randomly selected individual from the population has the characteristic of interest.

Since the sample proportion is defined to be the sample mean of a specific type of population (Bernoulli population), it inherits the same properties as the sample mean. We can use this fact to calculate the mean and variance of $\hat{p}$ as follows:

$$E(\hat{p}) = E(X_i) = p \quad \text{and} \quad \text{Var}(\hat{p}) = \frac{\text{Var}(X_i)}{n} = \frac{p(1-p)}{n}.$$

And surely, the CLT can also be used for $\hat{p}$ for a large sample size ($n \geq 30$):

$$\hat{p} \sim N\left(\mu = p, \sigma = \sqrt{\frac{p(1-p)}{n}}\right)$$

or in other word

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1).$$

**Example** A manufacturer claimed that $80\%$ of there manufactured DC motors have rotation speed in acceptable range. If we sample 100 DC motors from this manufacturer, what is the probability that no more than that $75\%$ of the sample have acceptable rotation speed?

We interested in the probability $P(\hat{p} \leq 0.75)$. For this, we can apply the CLT as follows:

$$
\begin{aligned}
P(\hat{p} \leq 0.75) &= P\left(\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{0.75-0.80}{\sqrt{(0.80)(0.20)/100}}\right) \\
&= P(Z \leq -1.25) \\
&= 0.10565
\end{aligned}
$$

---

## Proof of the Central Limit Theorem (OPTIONAL)

The standard tool for proving the central limit theorem is moment-generating function because it uniquely characterizes probability distributions and simplifies the process of handling sums of random variables.

Let $X_1, X_2, \ldots, X_n$ be a sample of size $n$ from an arbitrary distribution with mean $\mu$ and variance $\sigma^2$ and let $\bar{X}$ be a sample mean of the sample given by

$$
\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} T_n
$$

To prove the CLT using MGFs, we show that the MGF of $\bar{X}$ converges to the MGF of a standard normal distribution $Z$, which is:

$$
\begin{aligned}
M_Z(t) = E[e^{tZ}] &= \int_{-\infty}^{\infty} e^{tz} \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz - z^2/2}\, dz \\
&\vdots \\
&= e^{t^2/2}.
\end{aligned}
$$

Namely, to show that as $n \to \infty$,

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

we needed to show that

$$
\lim_{n \to \infty} M_{Z_{\bar{X}}}(t) = e^{t^2/2}.
$$

First, to simplify the notation, we let $Z_i = (X_i - \mu)/\sigma$ for $i = 1, 2, \ldots, n$ and $T_n = Z_1 + Z_2 + \cdots Z_n$ such that

$$
\frac{\bar{X}-\mu}{\sigma/\sqrt{(n)}} = \frac{T_n}{\sqrt{n}}.
$$

The result above can be proof as follows:

$$\begin{aligned}\frac{T_n}{\sqrt{n}} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma} = \frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}X_i-\mu \\ &= \frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^{n}X_i-n\mu\right) \\ &= \frac{1}{\sigma\sqrt{n}}\left(n\bar{X}-n\mu\right) = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\end{aligned}$$

Thus, we can write $Z_{\bar{X}} = T_n/\sqrt{n}$ and $M_{Z_{\bar{X}}}(t) = M_{T_n/\sqrt{n}}(t)$. This step is not that necessary but provide a short and concise notation we can use throughout the proof.

Now, with the new notation, we want to show that

$$\lim_{n\to\infty}\frac{T_n}{\sqrt{n}} \sim N(0,1) \quad \Leftrightarrow \quad \lim_{n\to\infty} M_{T_n/\sqrt{n}}(t) = e^{t^2/2}.$$

Consider the MGF of $T_n/\sqrt{n}$ as given by

$$\begin{aligned}M_{T_n/\sqrt{n}}(t) &= E\left[e^{t\cdot T_n/\sqrt{n}}\right] \\ &= E\left[e^{t(Z_1+Z_2+\cdots Z_n)/\sqrt{n}}\right] \\ &= E\left[e^{tZ_1/\sqrt{n}}e^{tZ_2/\sqrt{n}}\cdots e^{tZ_n/\sqrt{n}}\right]\end{aligned}$$

Since $X_i$'s are independent, and so does $Z_i$'s, so we can separate the expected value as follows.

$$M_{T_n/\sqrt{n}}(t) = E\left[e^{tZ_1/\sqrt{n}}\right]E\left[e^{tZ_2/\sqrt{n}}\right]\cdots E\left[e^{tZ_n/\sqrt{n}}\right]$$

The term $E[e^{tZ_i/\sqrt{n}}]$ is equal to the MGF of $Z_i$ but evaluated at $t/\sqrt{n}$ instead of just $t$, hence,

$$M_{T_n/\sqrt{n}}(t) = M_{Z_1}\left(\frac{t}{\sqrt{n}}\right)M_{Z_2}\left(\frac{t}{\sqrt{n}}\right)\cdots M_{Z_n}\left(\frac{t}{\sqrt{n}}\right)$$

Now, since $X_i$'s are identically distributed, again, and so does $Z_i$'s, so we can write

$$M_{T_n/\sqrt{n}}(t) = \left[M_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

And now, we evaluate this equation as $n$ goes to infinity:

$$\lim_{n\to\infty} M_{T_n/\sqrt{n}}(t) = \lim_{n\to\infty}\left[M_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n \to 1^\infty$$

Using the series expansion, we can easily observe that $M_Z(0) = 1$, which arise an indeterminate form, so we proceed to evaluate this limit using some algebraic manipulations and L'Hopital rule:

$$\begin{aligned}
\lim_{n\to\infty} M_{T_n/\sqrt{n}}(t) &= \lim_{n\to\infty}\left[M_Z\left(\frac{t}{\sqrt{n}}\right)\right]^n && \to 1^\infty \\
\ln\left(\lim_{n\to\infty} M_{T_n/\sqrt{n}}(t)\right) &= \lim_{n\to\infty} n\cdot\ln\left(M_Z(t/\sqrt{n})\right) && \to \infty\cdot 0 \\
&= \lim_{n\to\infty}\frac{\ln\left(M_Z(t/\sqrt{n})\right)}{n^{-1}} && \to \frac{0}{0} \\
&\overset{\text{L'H}}{=} \lim_{n\to\infty}\frac{M_Z(t/\sqrt{n})^{-1}\cdot M_Z'(t/\sqrt{n})\cdot\frac{-1}{2}tn^{-3/2}}{-n^{-2}} \\
&= \frac{1}{2}\lim_{n\to\infty}\frac{M_Z'(t/\sqrt{n})\cdot tn^{1/2}}{M_Z(t/\sqrt{n})} \\
&= \frac{1}{2M_Z(0)}\lim_{n\to\infty} M_Z'(t/\sqrt{n})\cdot tn^{1/2} \\
&= \frac{1}{2}\lim_{n\to\infty} M_Z'(t/\sqrt{n})\cdot tn^{1/2}
\end{aligned}$$

The term $\lim_{n\to\infty} M_Z'(t/\sqrt{n})$ or simply $M_Z'(0)$ is the first moment or the expected value of $Z_i$'s, which calculated to be:

$$\begin{aligned}
M_Z'(0) = E(Z_i) = E\left(\frac{X_i-\mu}{\sigma}\right) &= \frac{1}{\sigma}E(X_i) - \frac{\mu}{\sigma} \\
&= \frac{\mu}{\sigma} - \frac{\mu}{\sigma} \\
&= 0.
\end{aligned}$$

With this expected value, we get

$$
\begin{aligned}
\ln\left(\lim_{n\to\infty} M_{T_n/\sqrt{n}}(t)\right) &= \frac{1}{2}\lim_{n\to\infty} M_Z'(t/\sqrt{n})\cdot tn^{1/2} && \to 0\cdot\infty \\
&= \frac{1}{2}\lim_{n\to\infty} \frac{t\cdot M_Z'(t/\sqrt{n})}{n^{-1/2}} && \to \frac{0}{0} \\
&\overset{\text{L'H}}{=} \frac{1}{2}\lim_{n\to\infty} \frac{t\cdot M_Z''(t/\sqrt{n})\cdot \frac{-1}{2}tn^{-3/2}}{-\frac{1}{2}n^{-3/2}} \\
&= \frac{1}{2}\lim_{n\to\infty} t^2\cdot M_Z''(t/\sqrt{n}) \\
&= \frac{1}{2}t^2\cdot M_Z''(0)
\end{aligned}
$$

The last step is to evaluate $M_Z''(0)$ or the second moment of $Z_i$'s:

$$
\begin{aligned}
M_Z''(0) = E(Z_i^2) = E\left[\left(\frac{X_i-\mu}{\sigma}\right)^2\right] &= E\left[\left(\frac{(X_i-\mu)^2}{\sigma^2}\right)^2\right] \\
&= E\left[\frac{\sigma^2}{\sigma^2}\right] = E(1) = 1.
\end{aligned}
$$

Therefore

$$
\ln\left(\lim_{n\to\infty} M_{T_n/\sqrt{n}}(t)\right) = \frac{1}{2}t^2 M_Z''(0) = \frac{1}{2}t^2
$$

that is

$$
\lim_{n\to\infty} M_{T_n/\sqrt{n}}(t) = \lim_{n\to\infty} M_{Z_{\bar{X}}}(t) = e^{t^2/2}.
$$

Thus, the standardized sample mean $(\bar{X}-\mu)/\sigma$ of a sample has the moment generating function $M_{Z_{\bar{X}}}(t) = e^{t^2/2}$ which is the same as the moment-generating function of the standard normal distribution for arbitrary distribution of the population as $n \to \infty$, and from the uniqueness theorem of moment-generating function, we can conclude that any sample mean is normally distributed when the sample size goes to infinity as stated by the Central Limit Theorem.

# Distribution of Linear Combination

We can generalize the sample mean $\bar{X}$ and sample total $T_n$ into a type of random variable called *linear combination*.

> Given a collection of $n$ rv's $X_1, X_2, \ldots, X_n$ (not necessary independent or identically distributed) and $n$ numerical constants $a_1, a_2, \ldots, a_n$, then the rv given by
>
> $$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i$$
>
> is called a **linear combination** of the $X_i$'s.

> **The Case of Normal Distribution**
>
> If $X_1, X_2, \ldots, X_n$ are independent, normally distributed rv's (with possibly different means and/or variances), then any linear combination of the $X_i$'s also has a normal distribution.

Notice that the sample mean $\bar{X}$ is a special case of linear combination of rv's with $a_i$'s $= 1/n$ and $X_i$'s being independent and identically distributed (i.i.d.). The same goes for sample total $T_n$ but with $a_i$'s $= 1$.

Now let's consider the expected value and variance of a linear combination:

> Let $X_1, X_2, \ldots, X_n$ have mean values $\mu_1, \mu_2, \ldots, \mu_n$ respectively, and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ respectively. If $Y$ is a linear combination of $X_i$'s, $i = 1, 2, \ldots, n$, then
>
> 1. Whether or not the $X_i$'s are independent
>
> $$\begin{aligned} E(Y) &= E(a_1X_1 + a_2X_2 + \cdots + a_nX_n) \\ &= a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) \\ &= a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i. \end{aligned}$$

2. If $X_i$'s are independent

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n)\\ &= a_1^2\text{Var}(X_1) + a_2^2\text{Var}(X_2) + \cdots + a_n^2\text{Var}(X_n)\\ &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2.\end{aligned}$$

3. For any $X_i$'s

$$\text{Var}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_ia_j\text{Cov}(X_i, X_j).$$

**Example** A gas station sells three grades of gasoline: regular, extra, and super. These are priced at $\$3.00$, $\$3.20$, and $\$3.40$ per gallon, respectively. Let $X_1$, $X_2$, and $X_3$ denote the amounts of these grades purchased (gallons) on a particular day. Suppose the $X_i$'s are independent with $\mu_1 = 1000, \mu_2 = 500, \mu_3 = 300, \sigma_1 = 100, \sigma_2 = 80,$ and $\sigma_3 = 50$. The revenue from sales is $Y = 3.0X_1 + 3.2X_2 + 3.4X_3$.

Thus, the mean, variance, and standard deviation of the revenue are

$$E(Y) = 3.0\mu_1 + 3.2\mu_2 + 3.4\mu_3 = \$5620$$

$$\text{Var}(Y) = (3.0)^2\sigma_1^2 + (3.2)^2\sigma_2^2 + (3.4)^2\sigma_3^2 = 184,436$$

$$\sigma_Y = \sqrt{1184,436} = \$429.46$$

## Distribution of Sample Variances

In addition to sample mean, one may want to study the variability of the population, in such case, the sampling distribution of sample variance $S^2$ can be used in learning about it parametric counterpart, the population variance $\sigma^2$.

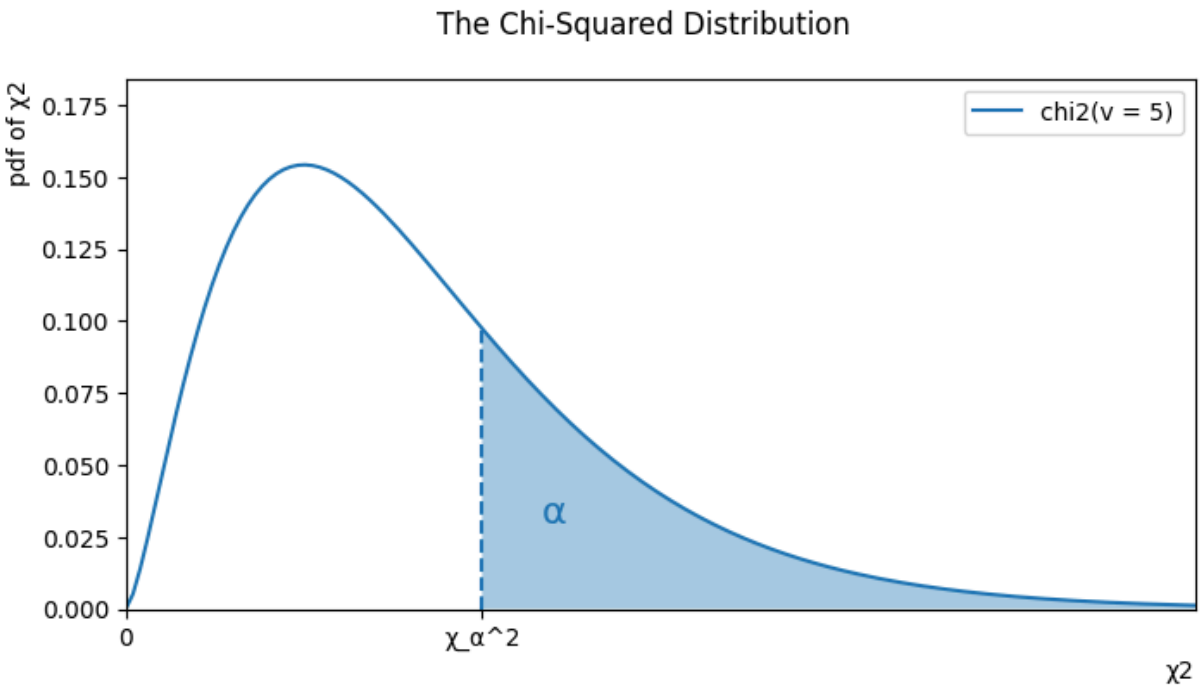
If $S^2$ is the variance of a random sample $X_1, X_2, \ldots, X_n$ of size $n$ taken from a **normal population** having the variance $\sigma^2$, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2}$$

has a **chi-squared distribution** with $v = n - 1$ degrees of freedom.

The probability that a random sample produces a $\chi^2$ value greater than some specified value is equal to the area under the curve to the right of this value. It is customary to let $\chi^2_\alpha$ represent the $\chi^2$ value above which we find an area of $\alpha$. The value of $\chi^2_\alpha$ is useful in analysis, so it often get tabulated for various degrees of freedom (For an example table, see Chi-Square Distribution Table .)

In [7]:
```
fig, ax = plt.subplots(1,1,figsize=(8,4))

fig.suptitle("The Chi-Squared Distribution")

xs = np.linspace(0,18,150)
ys = chi2.pdf(xs, 5)

xa = 6
xsa = np.linspace(6,18,150)
ysa = chi2.pdf(xsa, 5)

ax.plot(xs, ys, label='chi2(v = 5)')
ax.plot([xa,xa], [0,chi2.pdf(xa,5)], color='tab:blue', linestyle='--')
ax.fill_between(
    xsa, xsa*0, ysa,
    alpha=0.4
)
ax.text(7, 0.03, 'α', fontsize=16, color='tab:blue')
ax.set(
    xlim=[0,18], ylim=[0,ys.max() + 0.03],
    xticks=[0,xa], xticklabels=['0', 'χ_α^2']
)
ax.set_xlabel('χ2', loc='right')
ax.set_ylabel('pdf of χ2', loc='top')
ax.legend();
```



Exactly 95% of a chi-squared distribution lies between $\chi^2_{0.975}$ and $\chi^2_{0.025}$. A $\chi^2$ value falling to the right of $\chi^2_{0.025}$ is not likely to occur unless our assumed value of $\sigma^2$ is too small. Similarly, a $\chi^2$ value falling to the left of $\chi^2_{0.975}$ is unlikely unless our assumed value of $\sigma^2$ is too large. In other words, it is possible to have a $\chi^2$ value to the left of $\chi^2_{0.975}$ or to the right of $\chi^2_{0.025}$ when $\sigma^2$ is correct, but if this should occur, it is more probable that the assumed value of $\sigma^2$ is in error.

**Example** A manufacturer of car batteries guarantees that the batteries will last, on average, $3$ years with a standard deviation of $1$ year. If five of these batteries have lifetimes of $1.9$, $2.4$, $3.0$, $3.5$, and $4.2$ years, should the manufacturer still be convinced that the batteries have a standard deviation of $1$ year? Assume that the battery lifetime follows a normal distribution.

We first find the sample variance:

$$s^2 = \frac{(1.9-3.0)^2 + (2.4-3.0)^2 + (3.0-3.0)^2 + (3.5-3.0)^2 + (4.2-3.0)^2}{5-1} = 0.815$$

And so

$$\chi^2 = \frac{(5-1)(0.815)}{1^2} = 3.26$$

is a value from a chi-squared distribution with $v = 4$ degrees of freedom. Since 95% of the $\chi^2$ values with $4$ degrees of freedom fall between $\chi^2_{0.975} = 0.484$ and $\chi^2_{0.025} = 11.143$, thus, the computed value $\chi^2 = 3.26 \in [0.484, 11.143]$ with $\sigma^2 = 1$ is reasonable, and therefore the manufacturer has no reason to suspect that the standard deviation is other than 1 year.

# t-Distribution

In the previous section, we discussed the used of the Central Limit Theorem in inferences on a population mean. However, the method assumed that the population standard deviation is known, as indicated by the formula containing the value $\sigma$:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

This assumption may not be unreasonable in situations where the engineer is quite familiar with the system or process. However, in many experimental scenarios, knowledge of $\sigma$ is certainly no more reasonable than knowledge of the population mean $\mu$. Often, in fact, an estimate of $\sigma$ must be supplied by the same sample information that produced the sample average $\bar{x}$. As a result, a natural statistic to consider to deal with inferences on $\mu$ is

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

If the sample size is small, the values of $S^2$ fluctuate considerably from sample to sample and the distribution of $T$ deviates substantially from that of a standard normal distribution of $Z$.

If the sample size is large enough, say $n \geq 30$, the distribution of $T$ does not differ considerably from the standard normal and we can considerably assume $T = Z$. However smaller size $n$ (e.g. $n < 30$), it is useful to deal with the exact distribution of $T$.

In developing the sampling distribution of $T$, we can then write

$$T = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{(S/\sqrt{n})/(\sigma/\sqrt{n})} = \frac{Z}{S/\sigma} = \frac{Z}{\sqrt{V/(n-1)}}$$

where

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad V = \frac{(n-1)S^2}{\sigma^2}$$

has a standard normal distribution and chi-squared distribution with $v = n - 1$ degrees of freedom respectively.

In sampling from *normal populations*, we can show that **$\bar{X}$ and $S^2$ are independent, and consequently so are $Z$ and $V$**. The following theorem gives the definition of a random variable $T$ as a function of $Z$ (standard normal) and $\chi^2$.

> Let $Z$ be a standard normal random variable and $V$ a chi-squared random variable with $v$ degrees of freedom. If Z and V are independent, then the distribution of the random variable $T$, where
>
> $$T = \frac{Z}{\sqrt{V/v}},$$
>
> is given by the density function
>
> $$h(t) = \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)\sqrt{\pi v}} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2}, \quad t \in \mathrm{R}.$$
>
> This is known as the ***t*-distribution** with $v$ degrees of freedom.

From the foregoing and the theorem above we have the following corollary.

> Let $X_1, X_2, \ldots, X_n$ be a random sample from normal population with population mean $\mu$ and population standard deviation $\sigma$. Let the sample mean and sample variance given by
>
> $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$
>
> Then the random variable (statistics) given by
>
> $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$
>
> has a $t$-distribution with $v = n - 1$ degrees of freedom.

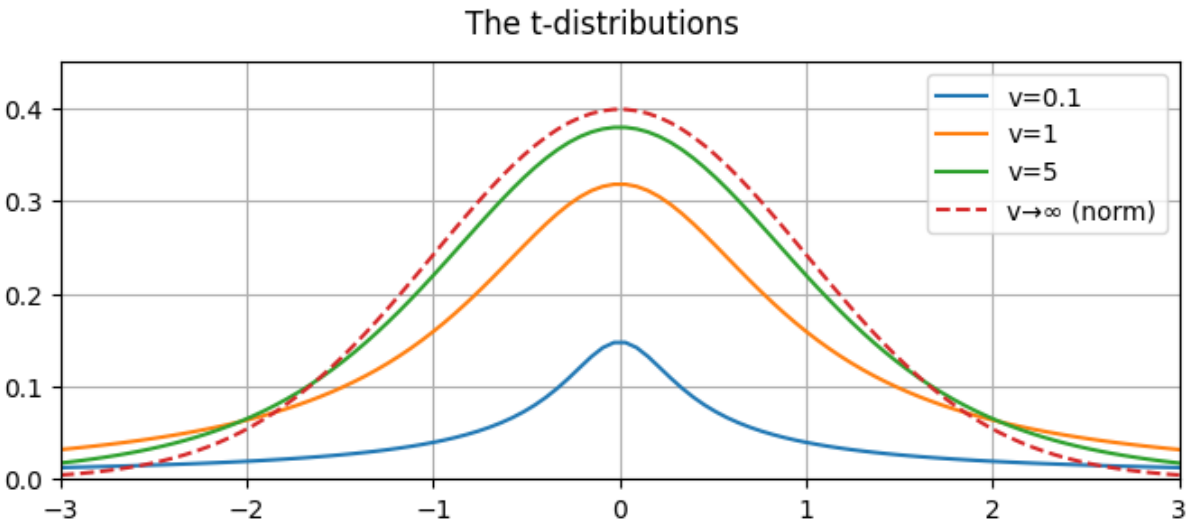
The statement assumed that the samples were selected from a normal population. Although this would seem to be a very restrictive assumption, it can be shown that non-normal populations possessing nearly bell-shaped distributions will still provide values of $T$ that approximate the $t$-distribution very closely.

Fun fact: The probability distribution of $T$ was first published in 1908 in a paper written by W. S. Gosset. At the time, Gosset was employed by an Irish brewery that prohibited publication of research by members of its staff. To circumvent this restriction, he published his work secretly under the name ***"Student."*** Consequently, the distribution of $T$ is usually called the **Student $t$-distribution** or simply the $t$-distribution.

The distribution of $T$ is similar to the distribution of $Z$ in that they both are symmetric about a mean of zero. Both distributions are bell shaped, but the $t$-distribution is more variable, owing to the fact that the $T$-values depend on the fluctuations of two quantities, $\bar{X}$ and $S^2$, whereas the $Z$-values depend only on the changes in $\bar{X}$ from sample to sample. Only when the sample size $n \to \infty$ (or in the other word, $v \to \infty$) will the two distributions become the same.

In [8]:
```
fig, ax = plt.subplots(1,1,figsize=(8,3))

fig.suptitle("The t-distributions")

xs = np.linspace(-3,3,100)

ax.plot(xs, student_t.pdf(xs, 0.1), label="v=0.1")
ax.plot(xs, student_t.pdf(xs, 1), label="v=1")
ax.plot(xs, student_t.pdf(xs, 5), label="v=5")
ax.plot(xs, student_t.pdf(xs, 999999999), label="v→∞ (norm)", linestyle="--")

ax.set(
    xlim=[-3,3], ylim=[0,0.45],
    axisbelow=True
)
```

```
ax.legend()
ax.grid();
```




Similar to chi-squared distribution, it is customary to let $t_\alpha$ represent the $t$-value above which we find an area equal to $\alpha$ (For an example table of values of $t_\alpha$, see Critical Values for Student's t-Distribution.)

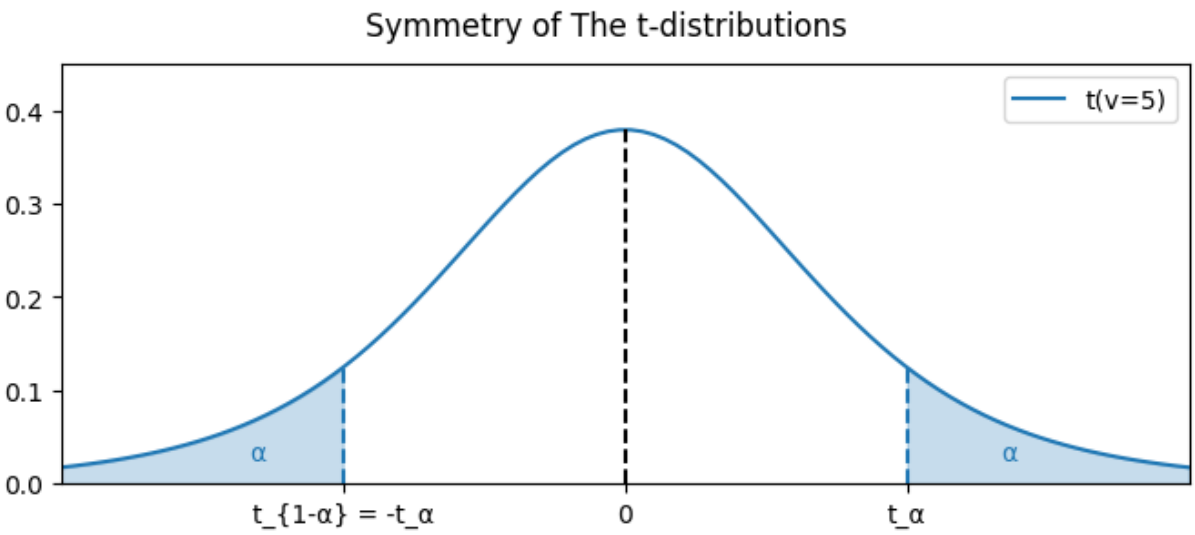
Since the $t$-distribution is symmetric about a mean of zero, we have

$$t_{1-\alpha} = -t_\alpha$$

that is, the $t$-value leaving an area of $1-\alpha$ to the right and therefore an area of $\alpha$ to the left is equal to the negative $t$-value that leaves an area of $\alpha$ in the right tail of the distribution. For example, $t_{0.95} = -t_{0.05}$, $t_{0.99} = -t_{0.01}$, and so forth.

In [9]:
```
fig, ax = plt.subplots(1,1,figsize=(8,3))

fig.suptitle("Symmetry of The t-distributions")

xs = np.linspace(-3,3,100)
ys = student_t.pdf(xs, 5)

ax.plot(xs, ys, label="t(v=5)")
ax.plot([0,0], [0,student_t.pdf(0, 5)], linestyle="--", color="black")
ax.plot([1.5,1.5], [0,student_t.pdf(1.5, 5)], linestyle="--", color="tab:blue")
ax.plot([-1.5,-1.5], [0,student_t.pdf(-1.5, 5)], linestyle="--", color="tab:blue")

xa = np.linspace(-3,-1.5,100)
xb = np.linspace(1.5,3,100)
ya = student_t.pdf(xa, 5)
yb = student_t.pdf(xb, 5)

ax.fill_between(xa, ya, ya*0, color="tab:blue", alpha=0.25)
ax.fill_between(xb, yb, yb*0, color="tab:blue", alpha=0.25)
ax.text(2, 0.025, "α", color="tab:blue")
ax.text(-2, 0.025, "α", color="tab:blue")
```

```
ax.set(
    xlim=[-3,3], ylim=[0,0.45],
    xticks=[-1.5,0,1.5], xticklabels=["t_{1-α} = -t_α", "0","t_α"]
)
ax.legend();
```




**Example** Find $k$ such that $P(k < T < -1.761) = 0.045$ for a random sample of size 15 selected from a normal distribution and

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

First Consider

$$\begin{aligned} P(k < T < -1.761) = 0.045 &= 1 - P(T < k) - P(T > -1.761) \\ &= 1 - (\text{area to the left of } k) - (\text{area to the right of } -1.761) \\ 0.995 &= (\text{area to the left of } k) + (\text{area to the right of } -1.761) \end{aligned}$$

From the t table, we know that for $t$-distribution with $v = 14$, $t_{0.05} = 1.761$. Hence

$$-t_{0.05} = -1.761 = t_{0.95}$$

so the area to the right of $-1.761$ or $P(T > -1.761)$ is 0.95, so we have

$$\begin{aligned} 0.995 &= (\text{area to the left of } k) + 0.95 \\ 0.005 &= (\text{area to the left of } k) \end{aligned}$$

that is $-t_{0.005} = k$ and therefore, from the table, $k = -2.977$.

---

## Usages of t-Distribution

The t-distribution is used extensively in problems that deal with inference about the population mean or in problems that involve comparative samples (i.e., in cases where one is trying to determine if means from two samples are significantly different).

**Example** A chemical engineer claims that the population mean yield of a certain batch process is $500$ grams per milliliter of raw material. To check this claim he samples $25$ batches each month. If the computed $t$-value falls between $-t_{0.05}$ and $t_{0.05}$ (90% probability interval), he is satisfied with this claim. What conclusion should he draw from a sample that has a mean $\bar{X} = 518$ grams per milliliter and a sample standard deviation $s = 40$ grams? Assume the distribution of yields to be approximately normal.

For a $t$-distribution with $24$ degrees of freedom, we have $t_{0.05} = 1.711$. So the satisfaction interval for the engineer is $t \in [-1.711, 1.711]$. With population mean of $\mu = 500$ we get

$$t = \frac{518 - 500}{40/\sqrt{25}} = 2.25,$$

which is above $1.711$, not in the satisfaction interval. The probability of seeing a $t$-value at least this much is

In [10]:
```
print("P(T > 2.25) =", 1 - student_t.cdf(2.25, 24))
```

```
P(T > 2.25) = 0.01694425545275391
```

which is not likely to happen based on current assumption. If $\mu > 500$, the value of $t$ computed from the sample is more reasonable. Hence, the engineer is likely to conclude that the process produces a better product (with more grams per milliliter of raw material) than he thought.

---

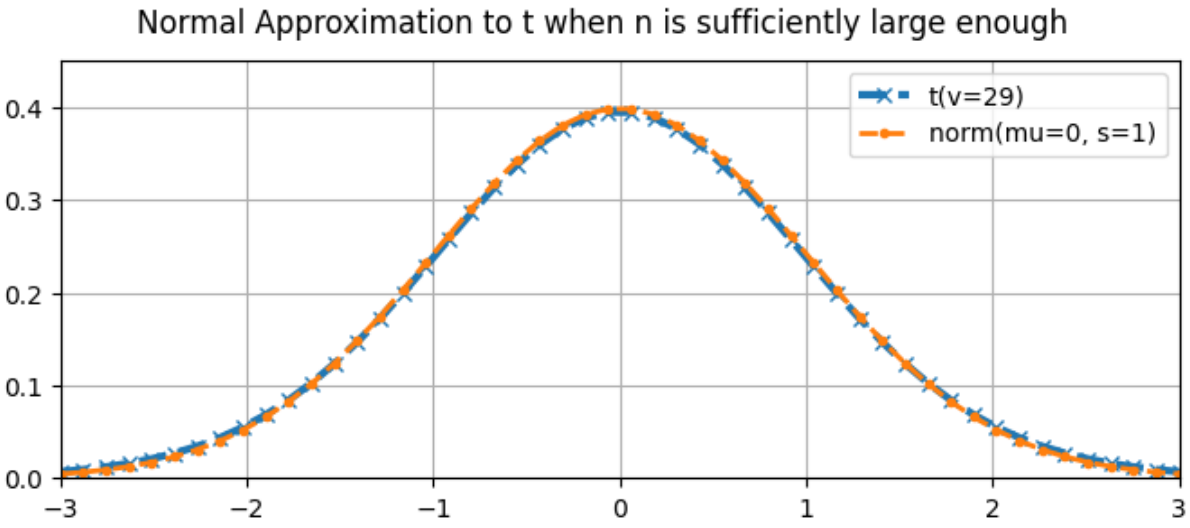
The use of the distribution will be extended in the following chapters. The reader should note that use of the $t$-distribution for the statistic

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

requires that $X_1, X_2, \ldots, X_n$ be **normal**. The use of the $t$-distribution and the sample size consideration do not relate to the Central Limit Theorem. The use of the standard normal distribution rather than $T$ for $n \geq 30$ merely implies that $S$ *is a sufficiently good estimator of* $\sigma$ in this case.

In [11]:
```
fig, ax = plt.subplots(1,1,figsize=(8,3))

fig.suptitle("Normal Approximation to t when n is sufficiently large enough")

xs = np.linspace(-3,3,50)

ax.plot(xs, student_t.pdf(xs, 29), "--x", linewidth=3, label="t(v=29)")
```

```
ax.plot(xs, norm.pdf(xs)         , "--.", linewidth=2, label="norm(mu=0, s=1)")

ax.set(
    xlim=[-3,3], ylim=[0,0.45],
    axisbelow=True
)
ax.legend()
ax.grid();
```




## F-Distribution

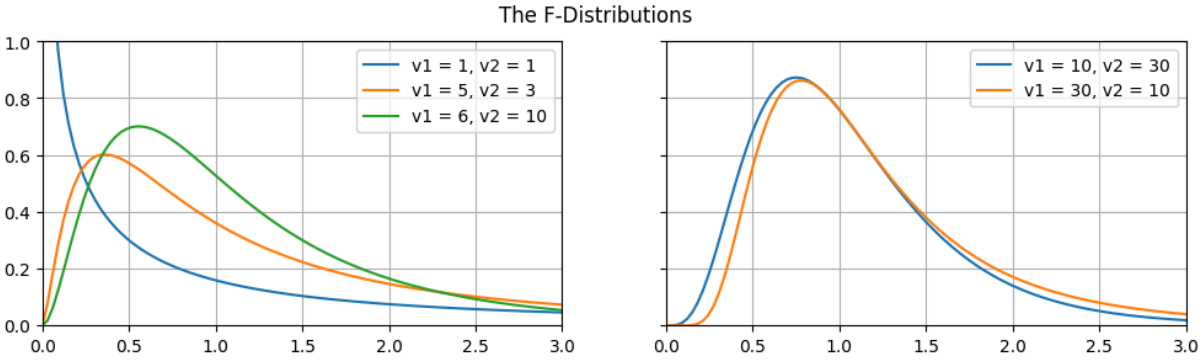
While it is of interest to let sample information shed light on two population means using $t$-distribution, it is often the case that a **comparison of variability** is equally important, if not more so. The $F$-distribution finds enormous application in comparing sample variances. Applications of the $F$-distribution are found in problems involving two or more samples.

> Let $U$ and $V$ be two independent random variables having *chi-squared distributions* with $v_1$ and $v_2$ degrees of freedom, respectively. Then the distribution of the random variable
>
> $$F_{(v_1,v_2)} = \frac{U/v_1}{V/v_2}$$
>
> is given by the density function
>
> $$h(f) = \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, \quad f > 0$$
>
> and $h(f) = 0$ for $f \leq 0$. This is known as the ***F*-distribution** with $v_1$ and $v_2$ degrees of freedom (d.f.).

# The density function will not be used and is given only for completeness.

The curve of the $F$-distribution depends not only on the two parameters $v_1$ and $v_2$ but also on the order in which we state them (swapping $v_1$ and $v_2$ results in different distribution curve). Do note that $v_1$ and $v_2$ are sometimes called degrees of freedom in numerator (d.f.N) and degrees of freedom in denominator (d.f.D.) respectively.

In [12]:
```
fig, axs = plt.subplots(1,2,figsize=(12,3),sharey=True)

fig.suptitle("The F-Distributions")

xs = np.linspace(0,3,100)

axs[0].plot(xs, f_dist.pdf(xs, 1, 1) , label="v1 = 1, v2 = 1")
axs[0].plot(xs, f_dist.pdf(xs, 5, 3) , label="v1 = 5, v2 = 3")
axs[0].plot(xs, f_dist.pdf(xs, 6, 10), label="v1 = 6, v2 = 10")

axs[1].plot(xs, f_dist.pdf(xs, 10, 30), label="v1 = 10, v2 = 30")
axs[1].plot(xs, f_dist.pdf(xs, 30, 10), label="v1 = 30, v2 = 10")

for ax in axs:
    ax.set(
        xlim=[0,3], ylim=[0,1],
        axisbelow=True
    )
    ax.legend()
    ax.grid();
```




Similar to previously discussed distributions, we can lat $f_\alpha$ be the $f$-value above which we find an area equal to $\alpha$ and these values often get tabulated (For an example table, see F Distribution Table.) The table often gives values of $f_\alpha$ only of some specific values of $\alpha$ of various values of $v_1$ and $v_2$. But by means of the following theorem, the table can also be used to find values like $f_{0.95}$ and $f_{0.99}$ from $f_{0.05}$ and $f_{0.01}$.

> Writing $f_\alpha(v_1, v_2)$ for $f_\alpha$ with $v_1$ and $v_2$ degrees of freedom, we obtain
>
> $$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

**Example** Find $k$ such that $P(F_{(5,3)} > k) = 0.95$.

From the definition of $f_{\alpha'}$ we can day that $k = f_{0.95}(5, 3)$. We can not find this value of the $F$-distribution table since it does not provide the $f_\alpha$ values of $\alpha = 0.95$. So we use the formula

$$k = f_{0.95}(5, 3) = \frac{1}{f_{0.05}(3, 5)}.$$

From the table we have $f_{0.05}(3, 5) = 5.41$ and so

$$k = f_{0.95}(5, 3) = \frac{1}{5.41} \approx 0.1848.$$

---

## The F-Distribution with Two Sample Variances

Suppose that random samples of size $n_1$ and $n_2$ are selected from two normal populations with variances $\sigma_1^2$ and $\sigma_2^2$, respectively. From previous section, we know that

$$\chi_1^2 = \frac{(n-1)S_1^2}{\sigma_1^2} \quad \text{and} \quad \chi_2^2 = \frac{(n-1)S_2^2}{\sigma_2^2}$$

are random variables having chi-squared distributions with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Furthermore, since the samples are selected at random, we are dealing with independent random variables. So we obtain the following result

> If $S_1^2$ and $S_2^2$ are the variances of independent random samples of size $n_1$ and $n_2$ taken from normal populations with variances $\sigma_1^2$ and $\sigma_2^2$, respectively, then
>
> $$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$
>
> has an $F$-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom.

**Example** A resin is mixed into an insulator material and subjected to an electric field to test the breakdown strength. One type of resin is mixed into $6$ insulators, and another type of resin is mixed into $9$ insulators. The electric field strength that causes breakdown is assumed to be normally distributed. The variance of the breakdown strength from the group using the first resin is $195$, and the variance from the group using the second resin is $215$. What is the probability that the ratio of the variances of breakdown strength for the sample group using the first resin to the second resin $(S_1^2/S_2^2)$ is greater than $6$.

We can simply calculate:

$$\begin{aligned} P\left(\frac{S_1^2}{S_2^2} > 6\right) &= P\left(\frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} > 6 \cdot \frac{\sigma_2^2}{\sigma_1^2}\right) \\ &= P\left(F_{(5,8)} > 6 \cdot \frac{215}{195}\right) \\ &= P(F_{(5,8)} > 6.6154) \end{aligned}$$

From the table, we can find the closest value to $6.6154$ to be $f_{0.01}(5, 8) = 6.63$. From this we can approximate that

$$P\left(\frac{S_1^2}{S_2^2} > 6\right) \approx f_{0.01}(5, 8) = 6.63$$

The more accurate value of probability is calculated to be:

In [13]:
```
print(1 - f_dist.cdf(6.6154, 5, 8), "(surprisingly accurate enough!)")
```

```
0.010074672162379783 (surprisingly accurate enough!)
```