Review of Statistical Theory
Part 3

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w

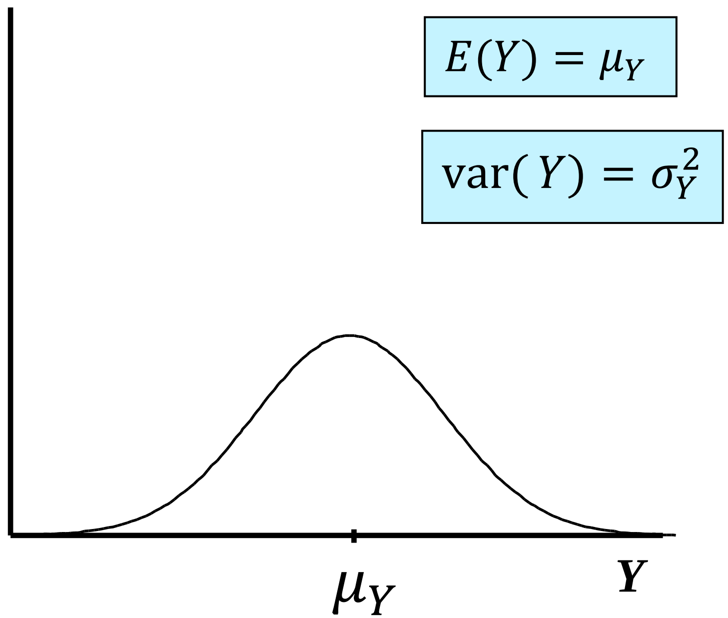
Brooklyn College

1. The probability framework for statistical inference
2. Estimation
3. **Hypothesis Testing**
4. Confidence intervals

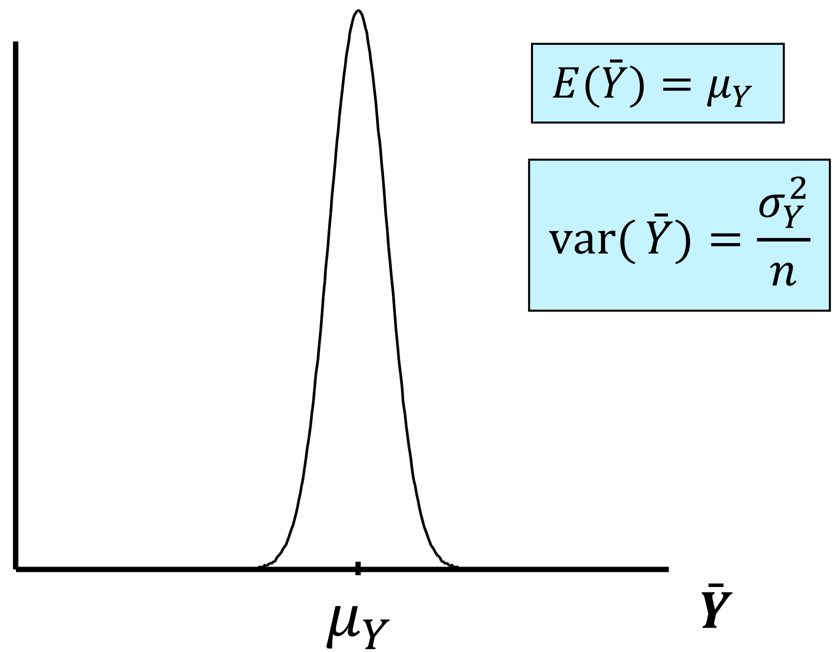
Distribution of the sample mean

In this illustration, Y and $\bar{Y}$ are both centered around μ_Y, but the dispersion differs.

Probability density function of Y



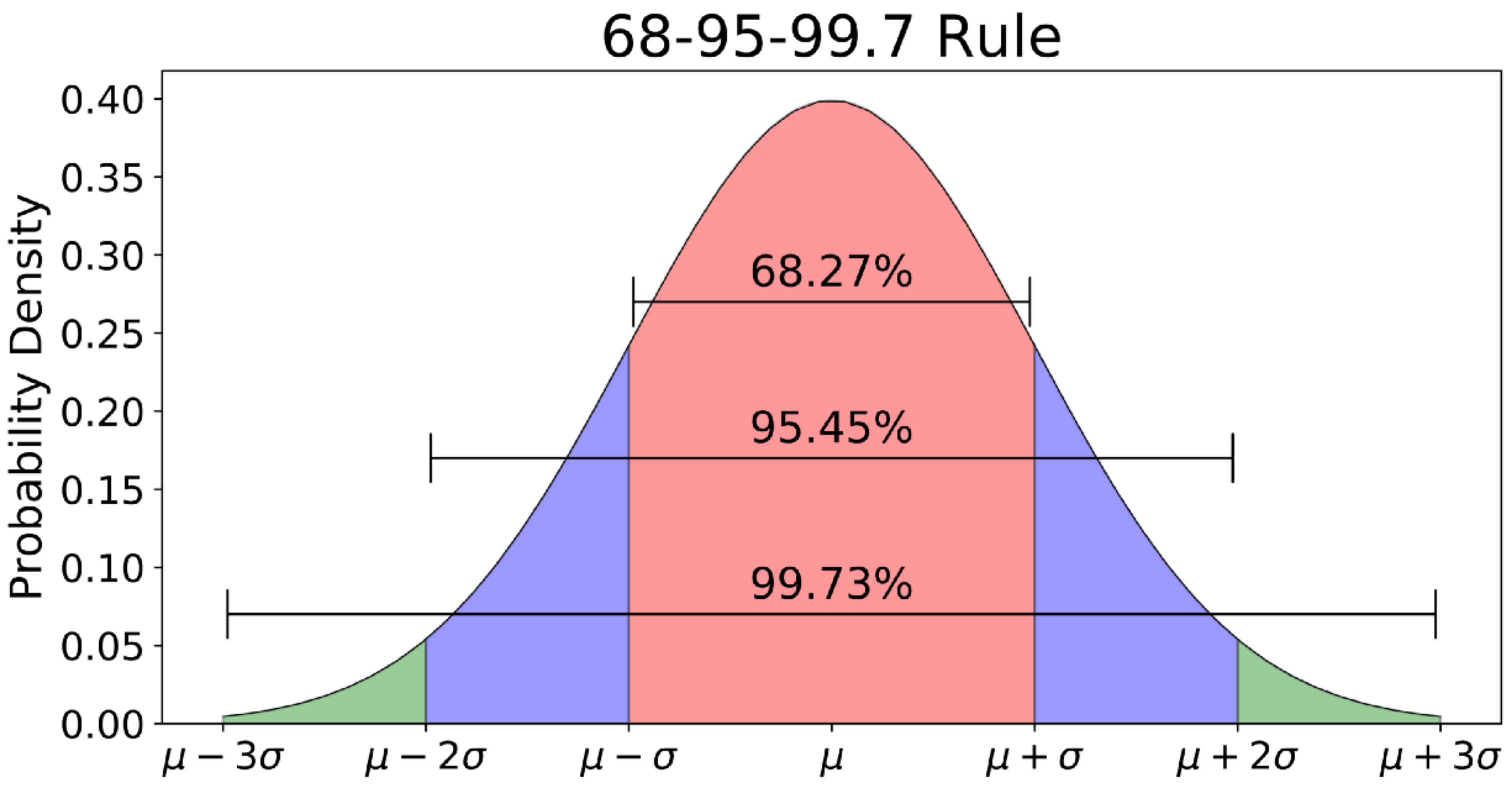
Probability density function of $\bar{Y}$



Note that $\bar{Y}$ is normally distributed even when the underlying random variable Y is not! Remember in our Central Limit Theorem (CLT) Excel simulation population Y was uniformly distributed.

https://wwwedu.github.io/BC4400S23/Lecture4/central_limit_theorem_simulation.xlsx

The normal distribution



Technically the rule should be 68.27 - 95.45 - 99.73 but 68 – 95 - 99 is easier to remember.

Exactly 95% of the distribution is found between $\mu - 1.96\sigma$ and $\mu + 1.96\sigma$

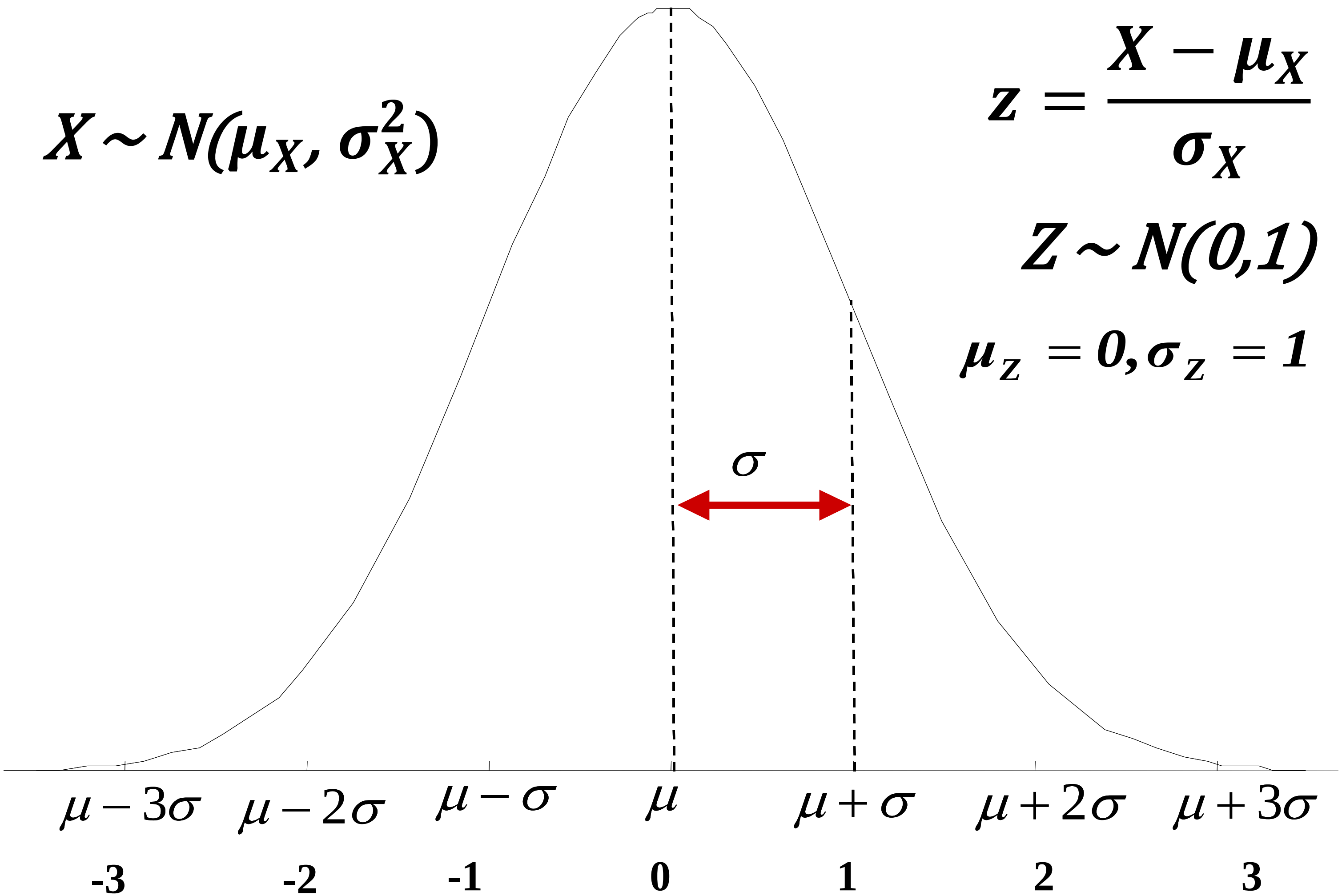
All normal distributions can be translated into standard normal

$$X \sim N(\mu_X, \sigma_X^2)$$

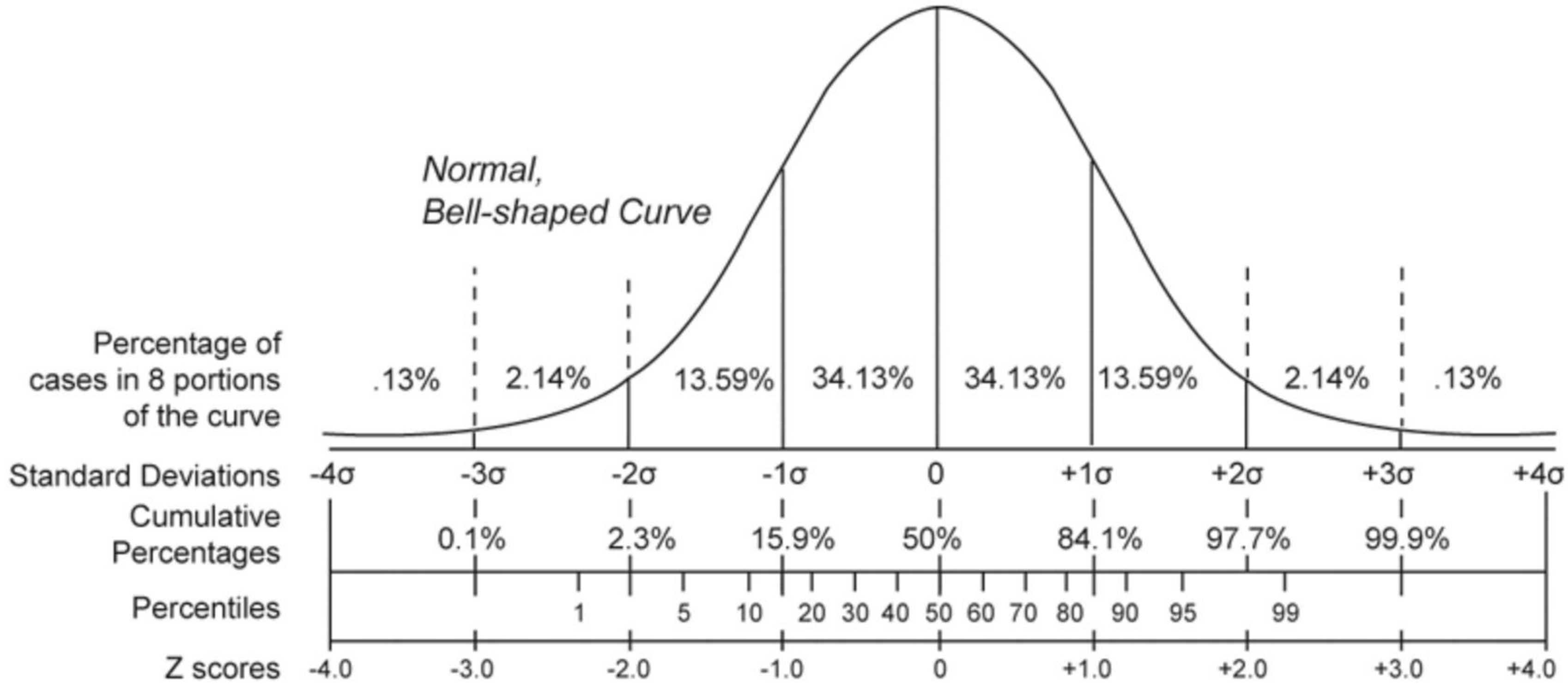
$$z = \frac{X - \mu_X}{\sigma_X}$$

$$Z \sim N(0,1)$$

$$\mu_Z = 0, \sigma_Z = 1$$



The standard normal profile



- For other values see statistical tables like the one in textbook Appendix: https://wwwedu.github.io/BC4400S23/Admin/StatsTables.pdf
- https://demonstrations.wolfram.com/AreaOfANormalDistribution/

Hypothesis Testing

The ***hypothesis testing*** problem (for the mean): make a provisional decision based on the evidence at hand whether a null hypothesis is true, or instead that some alternative hypothesis is true. That is, test

- H_0: $E(Y) \leq \mu_{Y,0}$ vs. H_1: $E(Y) > \mu_{Y,0}$ (1-sided, >)
- H_0: $E(Y) \geq \mu_{Y,0}$ vs. H_1: $E(Y) < \mu_{Y,0}$ (1-sided, <)
- H_0: $E(Y) = \mu_{Y,0}$ vs. H_1: $E(Y) \neq \mu_{Y,0}$ (2-sided)

Some terminology for testing statistical hypotheses (1 of 2)

p-value = probability of drawing a statistic (e.g. $\bar{Y}$) at least as adverse to the null as the value actually computed with your data, assuming that the null hypothesis is true.

The ***significance level*** of a test is a pre-specified probability of incorrectly rejecting the null, when the null is true.

Calculating the p-value based on $\bar{Y}$:

$$p - \text{value} = \boxed{\Pr[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|]}$$

Where $\bar{Y}^{act}$ is the value of $\bar{Y}$ actually observed (nonrandom)

Some terminology for testing statistical hypotheses (2 of 2)

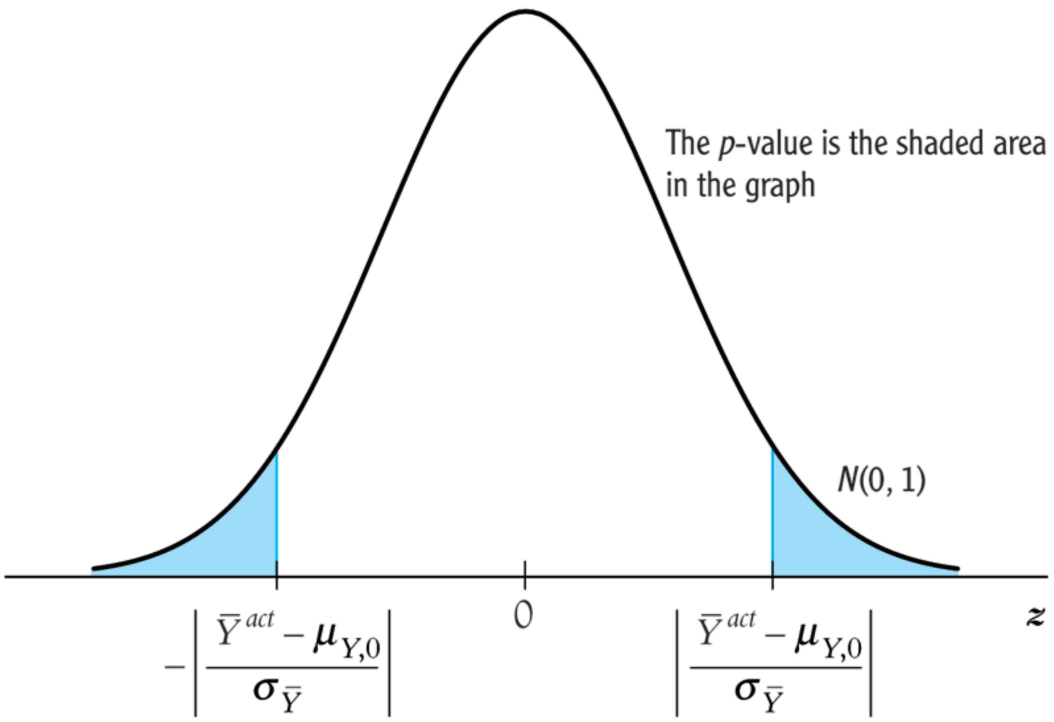
- To compute the *p*-value, you need the to know the sampling distribution of $\bar{Y}$, which is complicated if *n* is small.
- If *n* is large, you can use the normal approximation (CLT):

$$
\begin{aligned}
p\text{-value} &= \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|], \\
&= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y / \sqrt{n}}|] \\
&= \Pr_{H_0}[|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}| > |\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}|]
\end{aligned}
$$

$\cong$ probability under left + right $N(0,1)$ tails

where $\sigma_{\bar{Y}} = \text{std. dev. of the distribution of } \bar{Y} = \sigma_Y/\sqrt{n}$.

Calculating the p-value with σ_Y known:



- For large *n*, *p*-value = the probability that a *N*(0,1) random variable falls outside $|(\bar{Y}^{act} - \mu_{Y,0})/\sigma_{\bar{Y}}|$
- In practice, $\sigma_{\bar{Y}}$ is unknown – it must be estimated

Estimator of the variance of Y:

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{“sample variance of } Y\text{”}$$

Fact:

If $(Y_1,\ldots,Y_n)$ are i.i.d. and $E(Y^4) < \infty$, then

$$s_Y^2 \xrightarrow{p} \sigma_Y^2$$

The Standard Error of $\overline{Y}$

KEY CONCEPT 3.4

The standard error of $\overline{Y}$ is an estimator of the standard deviation of $\overline{Y}$. The standard error of $\overline{Y}$ is denoted $SE(\overline{Y})$ or $\hat{\sigma}_{\overline{Y}}$. When $Y_1, \ldots, Y_n$ are i.i.d.,

$$SE(\overline{Y}) = \hat{\sigma}_{\overline{Y}} = s_Y/\sqrt{n}. \tag{3.8}$$

Computing the p-value with σ_Y^2 estimated:

$$
\begin{aligned}
p\text{-value} &= \Pr_{H_0}[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|], \\
&= \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_Y/\sqrt{n}}\right|\right] \\
&\cong \Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right] \quad \text{(large } n\text{)}
\end{aligned}
$$

so

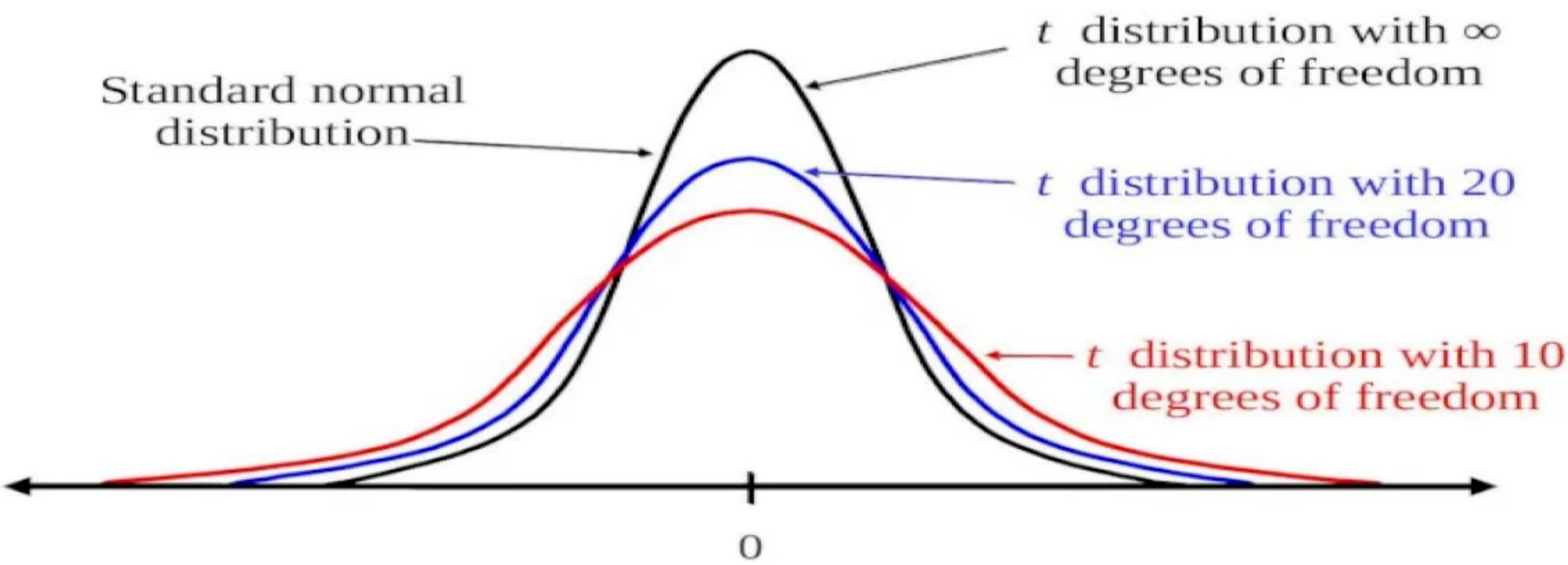
$$\boxed{p\text{-value} = \Pr_{H_0}[|t| > |t^{act}|]} \qquad (\sigma_Y^2 \text{ estimated})$$

$\cong$ probability under normal tails outside $|t^{act}|$

where $t = \dfrac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}$ (the usual $t-$statistic)

The Student-t distribution

The t-distribution is used when n is **small** and σ is **unknown**.



Compared to the standard normal the t-distribution has "fatter tails" (it is leptokurtic). That means "rare" events, those further from the mean, happen with a higher probability relative to the standard normal.

The Student-t distribution

TABLE 2 Critical Values for Two-Sided and One-Sided Tests Using the Student *t* Distribution

	Significance Level				
Degrees of Freedom	20% (2-Sided) 10% (1-Sided)	10% (2-Sided) 5% (1-Sided)	5% (2-Sided) 2.5% (1-Sided)	2% (2-Sided) 1% (1-Sided)	1% (2-Sided) 0.5% (1-Sided)
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

Note that as the degrees of freedom increases (larger sample sizes) we find 95% of the distribution between -1.96 and +1.96 standard errors around the mean, same as the standard normal distribution. But with smaller sample sizes, we have to go further away from the mean to include 95%.

The Student-t distribution

- For $n > 30$, the t-distribution and $N(0,1)$ are very close (as n grows large, the t_{n-1} distribution converges to $N(0,1)$)
- The t-distribution is an artifact from days when sample sizes were small and "computers" were people
- For historical reasons, statistical software typically uses the t-distribution to compute p-values – but this is irrelevant when the sample size is moderate or large.
- For these reasons, in this class we will focus on the large-n approximation given by the CLT

The t-test of significance: decision rules

Type of hypothesis	H_0 : the null hypothesis	H_1 : the alternative hypothesis	Decision rule: reject H_0 if
Two-tail	$\mu_Y = \mu_{Y,0}$	$\mu_Y \neq \mu_{Y,0}$	$\lvert t\rvert > t_{\alpha/2,\,\mathrm{df}}$
Right-tail	$\mu_Y \leq \mu_{Y,0}$	$\mu_Y > \mu_{Y,0}$	$t > t_{\alpha,\,\mathrm{df}}$
Left-tail	$\mu_Y \geq \mu_{Y,0}$	$\mu_Y < \mu_{Y,0}$	$t < -t_{\alpha,\,\mathrm{df}}$

Notes :

- $\mu_{Y,0}$ is the hypothesized numerical value of μ_Y.
- $|t|$ means the absolute value of t.
- $t_{\alpha,\,\mathrm{df}}$ or $t_{\alpha/2,\,\mathrm{df}}$ means the critical t value at the α or $\alpha/2$ level of significance.
- df: degrees of freedom, (n - 1) for the one parameter model

Hypothesis test example 1

The average adult male height in a certain country is 170 cm. We suspect that the men in a certain city in that country might have a different average height due to some environmental factors. We pick a random sample of size 9 from the adult males in the city and obtain the following values for their heights (in cm):

$$176.2 \quad 157.9 \quad 160.1 \quad 180.9 \quad 165.1 \quad 167.2 \quad 162.9 \quad 155.7 \quad 166.2$$

Assume that the height distribution in this population is normally distributed. Here, we need to decide between

$$H_0: \mu = 170$$
$$H_1: \mu \neq 170$$

Based on the observed data, is there enough evidence to reject H_0 at significance level $\alpha = 0.05$?

Hypothesis test example 1

Solution:

Let's first compute the sample mean and the sample standard deviation. The sample mean is

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9}{9} = 165.8$$

The sample variance is given by

$$S^2 = \frac{1}{9-1}\sum_{k=1}^{9}(X_k - \bar{X})^2 = 68.01$$

The sample standard deviation is given by $S = \sqrt{S^2} = 8.25$

Hypothesis test example 1

Now, our test statistic is

$$W(X_1, X_2, \cdots, X_9) = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} = \frac{165.8 - 170}{\frac{8.25}{3}} = -1.52$$

Thus, $|W| = 1.52$. Because we estimated σ_Y^2 using a small sample size, we have to use a t-distribution test. Checking the statistical tables, the critical value (the limit where we include 95% of the sample around the mean) with 9 degrees of freedom is:

$$t_{\frac{\alpha}{2}, n-1} = t_{0.025, 8} \approx 2.31$$

Thus, we conclude

$$|W| \leq t_{\frac{\alpha}{2}, n-1}$$

Therefore, we cannot reject H_0. In other words, we do not have enough evidence to conclude that the average height in the city is different from the average height in the country.

Hypothesis test example 2

Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of $n = 100$ students from one large high school had a mean score of 58. Is there evidence to suggest that this high school is inferior?

Hint: calculate the probability that the sample mean is at most 58 when $n = 100$.

Hypothesis test example 2

Let $\bar{X}$ denote the mean of a random sample of $n = 100$ scores from a population with $\mu = 60$ and $\sigma^2 = 64$. We want to approximate $P(\bar{X} \le 58)$. We know from the Central Limit Theorem that $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ has a distribution that can be approximated by a standard normal distribution.

$$P(\bar{X} \le 58) = P\left(\frac{\bar{X} - 60}{8/\sqrt{100}} \le \frac{58 - 60}{.8}\right) \approx P(Z \le -2.5) = .0062$$

Where 0.0062 is the p-value we find in the standard normal table for a z-score of -2.5. Because this probability is so small, it is unlikely that the sample from the school of interest can be regarded as a random sample from a population with $\mu = 60$ and $\sigma^2 = 64$. The evidence suggests that the average score for this high school is lower than the overall average of $\mu = 60$.

1. The probability framework for statistical inference
2. Estimation
3. Testing
4. **Confidence intervals**

Confidence Intervals

- A 95% ***confidence interval*** for μ_Y is an interval that contains the true value of μ_Y in 95% of repeated samples.
- *Digression*: What is random here? The values of $Y_1,\dots,Y_n$ and thus any functions of them – including the confidence interval. The confidence interval will differ from one sample to the next. The population parameter, μ_Y, is not random; we just don't know it.

Confidence Intervals

A 95% confidence interval can always be constructed as the set of values of μ_Y not rejected by a hypothesis test with a 5% significance level.

$$\{\mu_Y\colon \left|\frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}}\right| \leq 1.96\} = \{\mu_Y\colon -1.96 \leq \frac{\bar{Y}-\mu_Y}{s_Y/\sqrt{n}} \leq 1.96\}$$

$$= \{\mu_Y\colon\ -1.96\frac{s_Y}{\sqrt{n}} \leq -\mu_Y \leq 1.96\ \frac{s_Y}{\sqrt{n}}\}$$

$$= \{\mu_Y\ \in (\bar{Y}-1.96\ \frac{s_Y}{\sqrt{n}},\ \bar{Y}+1.96\ \frac{s_Y}{\sqrt{n}})\}$$

This confidence interval relies on the large-n results that $\bar{Y}$ is approximately normally distributed and $s_Y^2 \xrightarrow{p} \sigma_Y^2$.

Confidence interval example

In a sample of $25, \bar{x} = 1.63$ and $s = 0.51$. Construct a 95 percent confidence interval for μ_x.

Solution:

2.064 is the 95% critical value from a t distribution with 24 degrees of freedom. Thus, the confidence interval is $1.63 \pm [\frac{2.064(0.51)}{\sqrt{25}}]$ or $[1.4195, 1.8405]$.

Summary:

From the two assumptions of:

1. simple random sampling of a population, that is, $\{Y_i, i = 1,\ldots,n\}$ are i.i.d.
2. $0 < E(Y^4) < \infty$

we developed, for large samples (large n):

- Theory of estimation (sampling distribution of $\bar{Y}$)
- Theory of hypothesis testing (large-n distribution of t-statistic and computation of the p-value)
- Theory of confidence intervals