

Review of Statistical Theory

Part 3

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w - Fall 2022

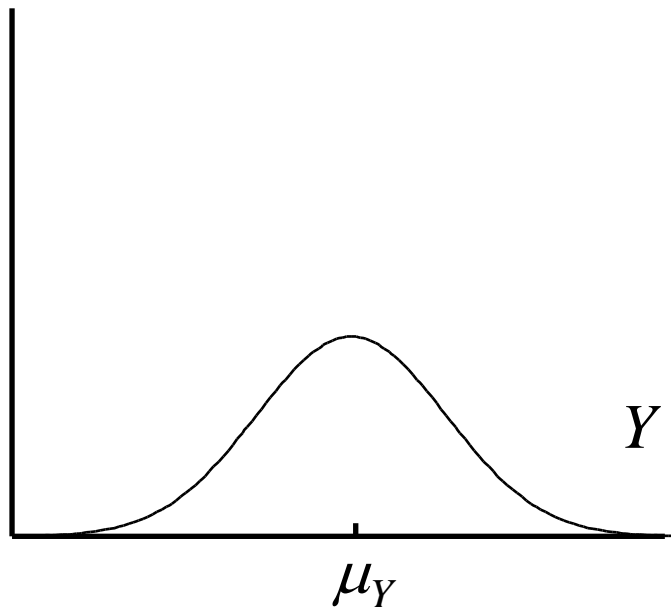
Brooklyn College
Sep 14, 2022

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

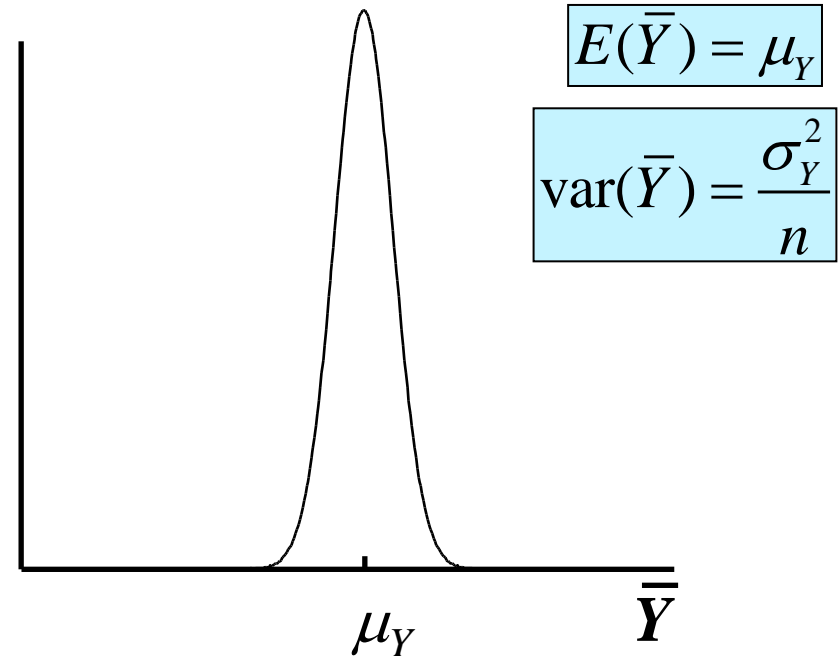
Distribution of the sample mean

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

probability density
function of Y

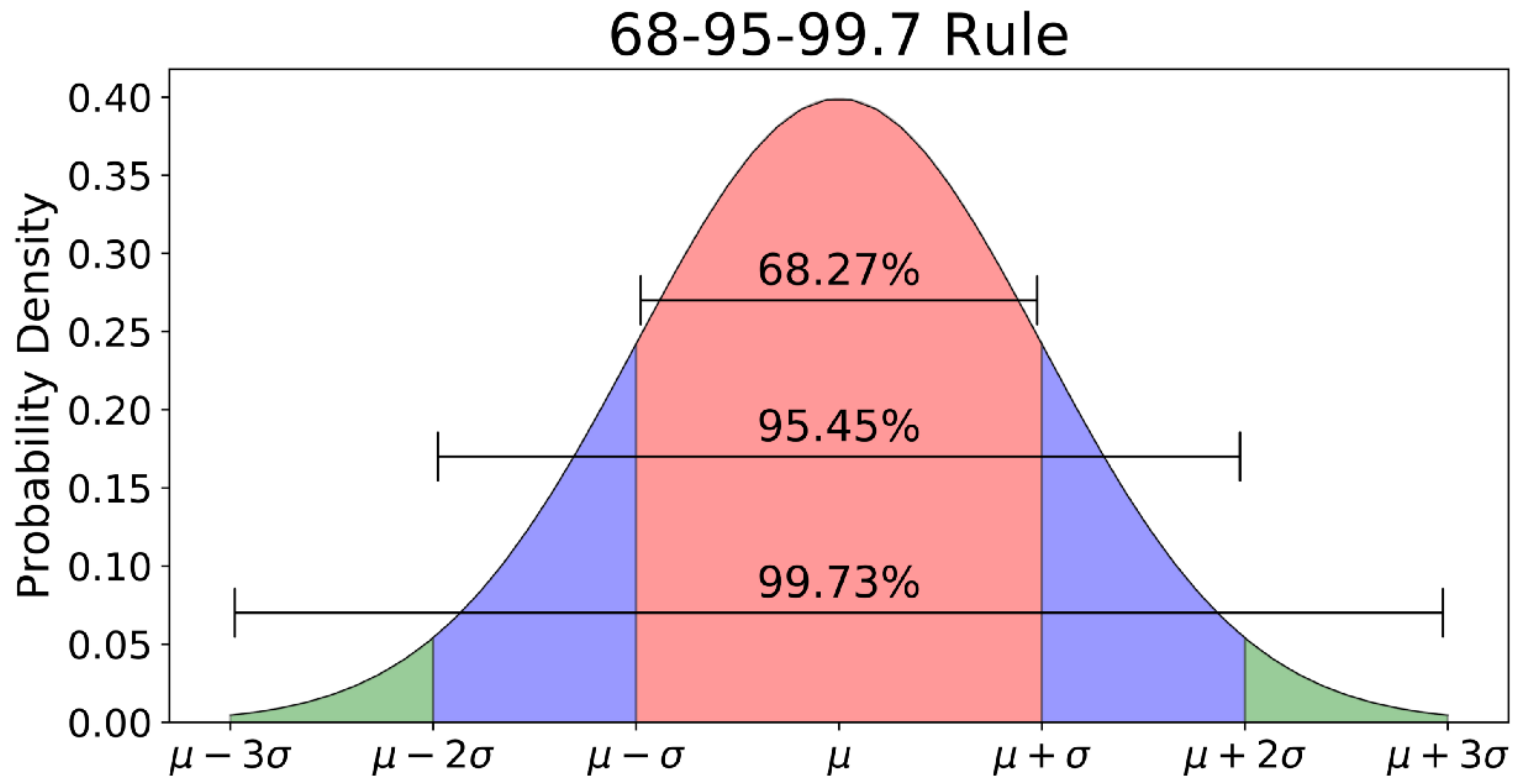


probability density
function of \bar{Y}

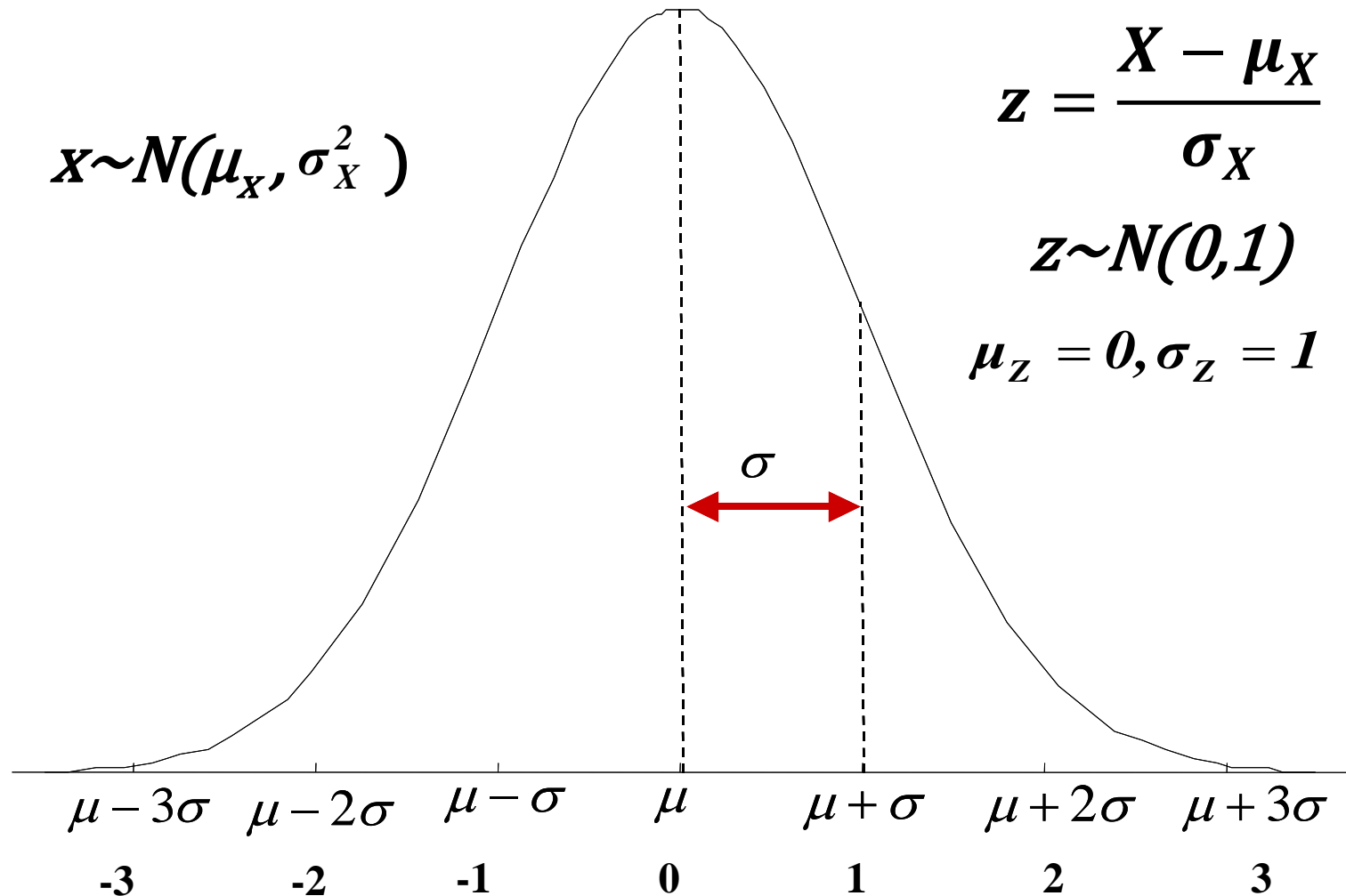


Note that \bar{Y} is normally distributed even when the underlying random variable Y is not! Remember in our CLT simulation Y was uniformly distributed.

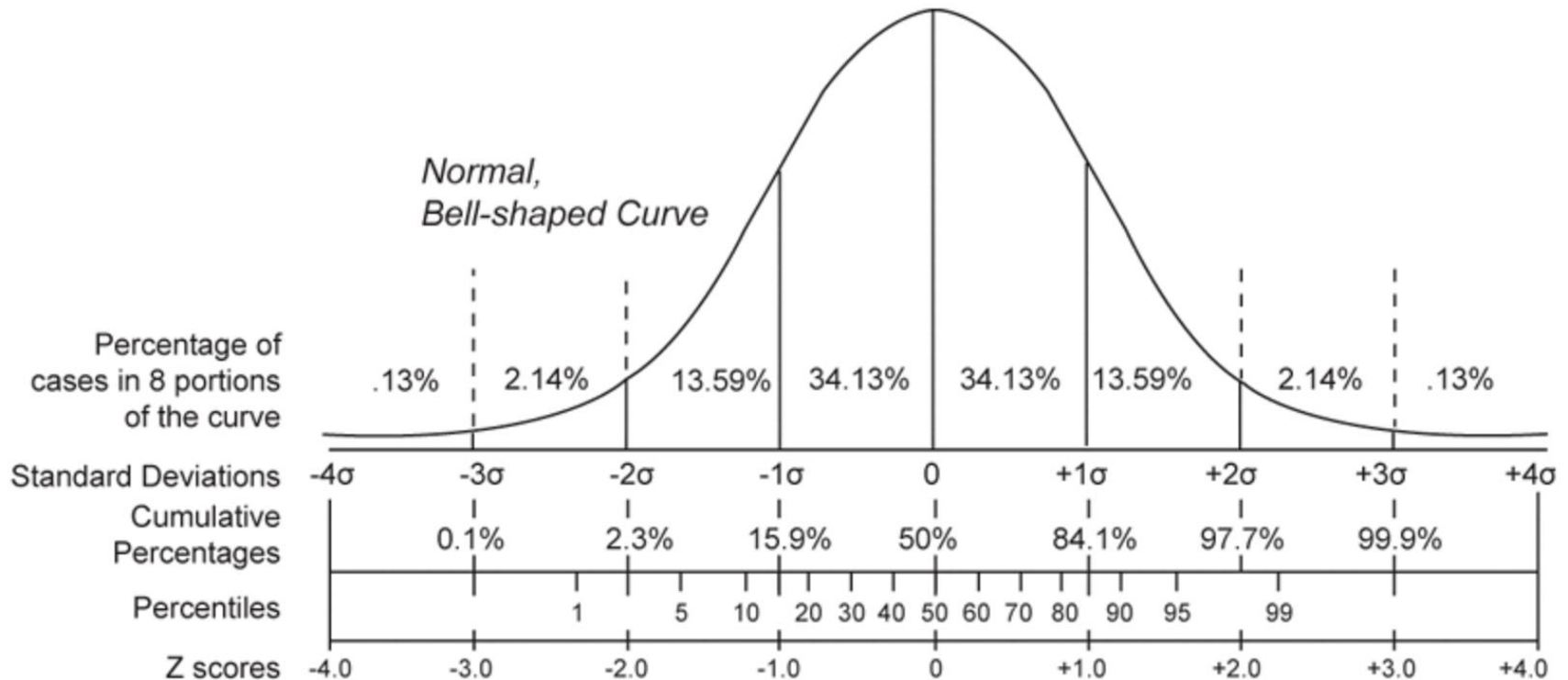
The normal distribution



All normal distributions can be translated into standard normal



The standard normal profile



- For other values see statistical tables like the one in textbook Appendix or here: <https://wwwedu.github.io/BC4400F22/Admin/StatsTables.pdf#page=2>
- <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

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

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

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

Some terminology for testing statistical hypotheses (2 of 2)

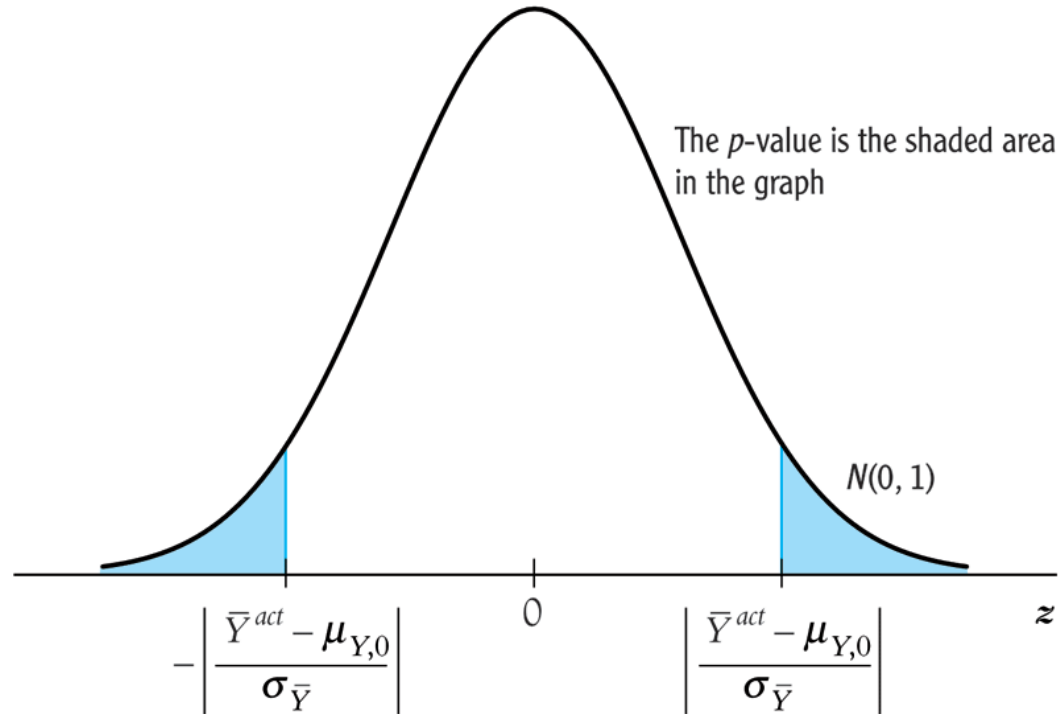
- To compute the p -value, you need to know the sampling distribution of \bar{Y} , which is complicated if n is small.
- If n is large, you can use the normal approximation (CLT):

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

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

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

The Standard Error of \bar{Y}

KEY CONCEPT

3.4

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

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

Computing the p -value with σ_Y^2 estimated:

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

so

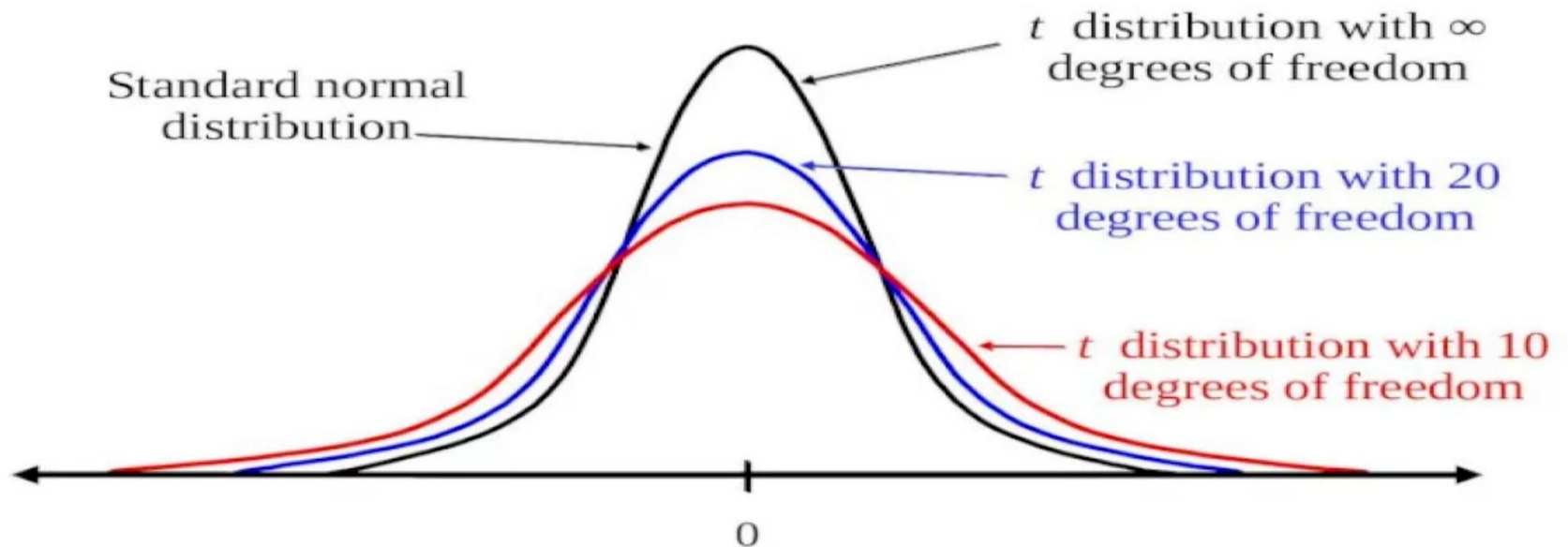
$$p\text{-value} = \Pr_{H_0} [|t| > |t^{act}|] \quad (\sigma_Y^2 \text{ estimated})$$

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

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

The Student-t distribution

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



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

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):

176.2 157.9 160.1 180.9 165.1 167.2 162.9 155.7 166.2

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

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

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

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

Thus, $|W| = 1.52$. Also, we have

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

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 $P(\bar{X} \leq 58)$. 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. Using the standard normal table we have:

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

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.

$$\begin{aligned}\{\mu_Y: \left| \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \right| \leq 1.96\} &= \{\mu_Y: -1.96 \leq \frac{\bar{Y} - \mu_Y}{s_Y / \sqrt{n}} \leq 1.96\} \\ &= \{\mu_Y: -1.96 \frac{s_Y}{\sqrt{n}} \leq -\mu_Y \leq 1.96 \frac{s_Y}{\sqrt{n}}\} \\ &= \{\mu_Y \in (\bar{Y} - 1.96 \frac{s_Y}{\sqrt{n}}, \bar{Y} + 1.96 \frac{s_Y}{\sqrt{n}})\}\end{aligned}$$

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 .

Solution:

2.064 is the 95% critical value from a t distribution with 24 degrees of freedom.

Thus, the confidence interval is $1.63 \pm [2.064(0.51)/5]$ or $[1.4195, 1.8405]$.

Summary:

From the two assumptions of:

1. simple random sampling of a population, that is, $\{Y_i, i = 1, \dots, n\}$ 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