

6. Statistical Inference and Hypothesis Testing

6.1 One Sample

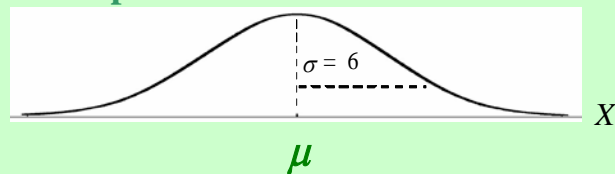
§ 6.1.1 Mean

STUDY POPULATION = Cancer patients on *new* drug treatment

Random Variable: X = “Survival time” (months)

Assume $X \approx N(\mu, \sigma)$, with unknown mean μ , but known (?) $\sigma = 6$ months.

Population Distribution of X



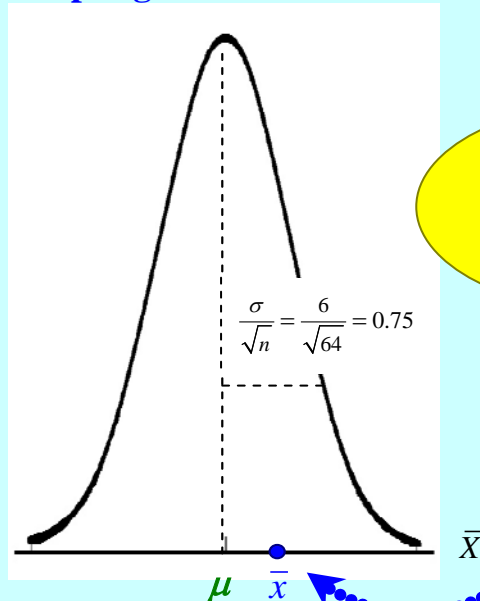
What can be said about the **mean μ** of this study population?



RANDOM SAMPLE, $n = 64$

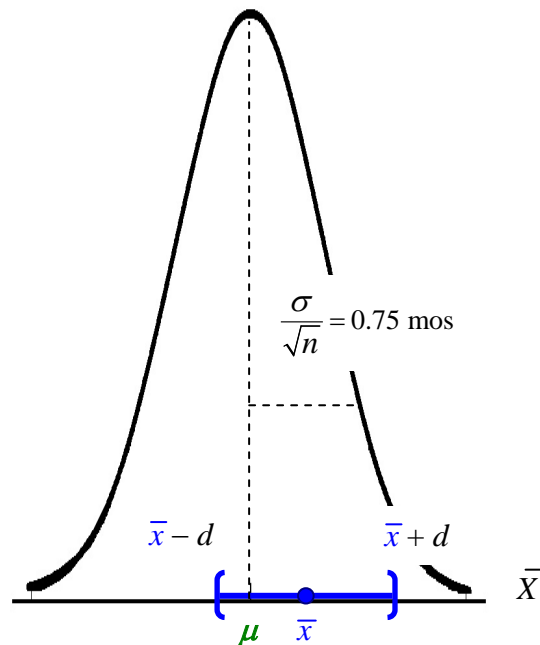
$\{x_1, x_2, x_3, x_4, x_5, \dots, x_{64}\}$

Sampling Distribution of \bar{X}



\bar{x} is called a
“point estimate”
of μ

Objective 1: Parameter Estimation ~ Calculate an **interval estimate** of μ , centered at the **point estimate** \bar{x} , that contains μ with a high probability, say 95%. (Hence, $1 - \alpha = 0.95$, so that $\alpha = 0.05$.)



That is, for any random sample, solve for d :

$$P(\bar{X} - d \leq \mu \leq \bar{X} + d) = 0.95$$

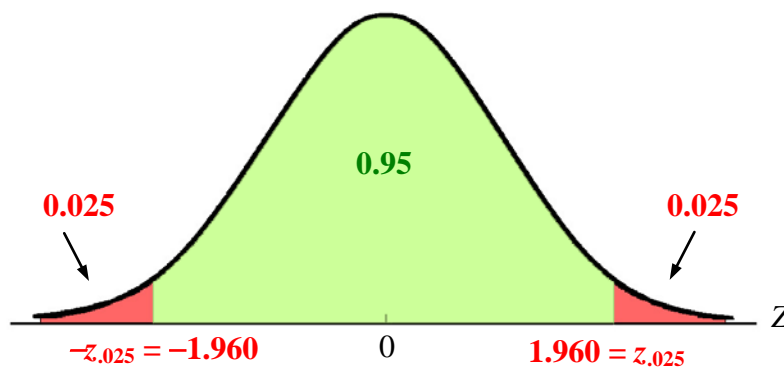
i.e., via some algebra,

$$P(\mu - d \leq \bar{X} \leq \mu + d) = 0.95.$$

For future reference, call this equation ★.

But recall that $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$. Therefore,

$$P\left(\frac{-d}{\sigma/\sqrt{n}} \leq Z \leq \frac{+d}{\sigma/\sqrt{n}}\right) = 0.95$$



Hence, $\frac{+d}{\sigma/\sqrt{n}} = z_{.025} \Rightarrow d = z_{.025} \times \frac{\sigma}{\sqrt{n}} = (1.96)(0.75 \text{ months}) = \mathbf{1.47 \text{ months.}}$
95% margin of error

95% Confidence Interval for μ

$$\left(\underbrace{\bar{x} - z_{.025} \frac{\sigma}{\sqrt{n}}}_{\text{Lower Limit}}, \underbrace{\bar{x} + z_{.025} \frac{\sigma}{\sqrt{n}}}_{\text{Upper Limit}} \right)$$

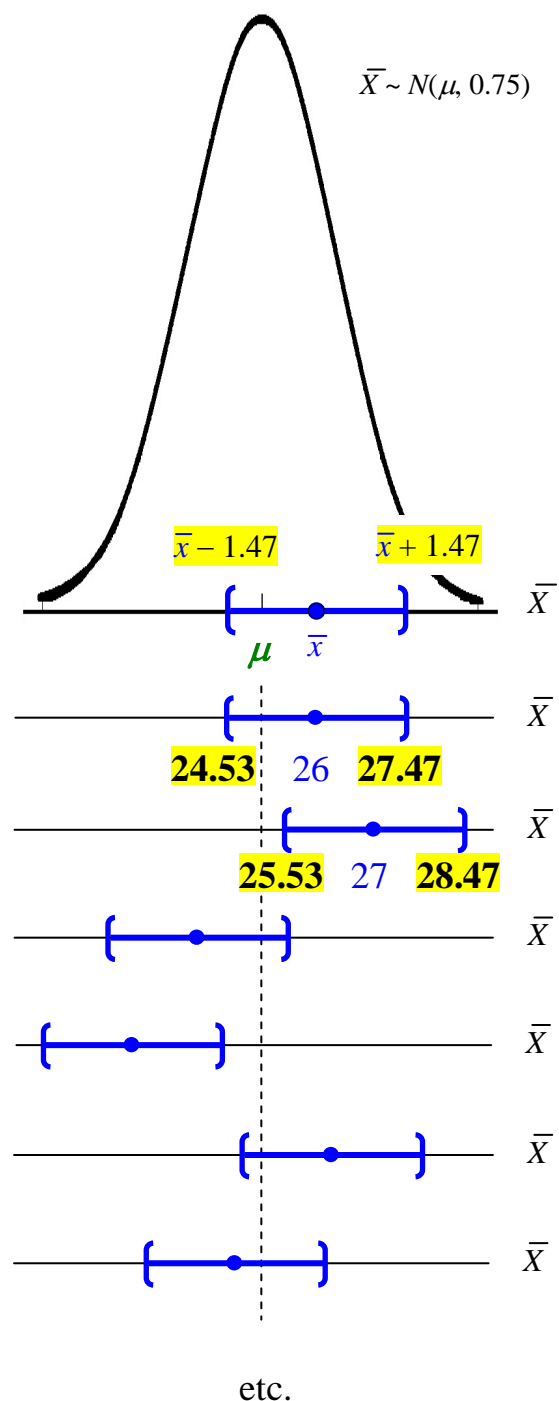
95% Confidence Limits

where the critical value $z_{.025} = 1.96$.

Therefore, the **margin of error** (and thus, the size of the confidence interval) remains the same, from sample to sample.

Example:

Sample	Mean \bar{x}	95% CI	
1	26.0 mos	$(26 - 1.47, 26 + 1.47)$	=
2	27.0 mos	$(27 - 1.47, 27 + 1.47)$	=
.	.	.	
.	.	.	
.	.	.	
.	.	.	
.	.	.	



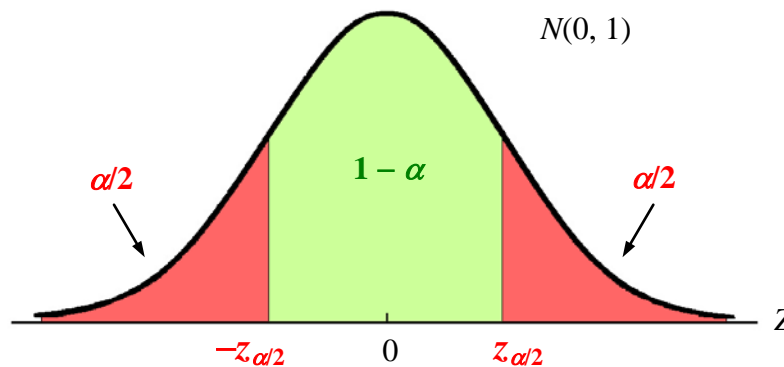
Interpretation: Based on Sample 1, the true mean μ of the “new treatment” population is between 24.53 and 27.47 months, with 95% “confidence.” Based on Sample 2, the true mean μ is between 25.53 and 28.47 months, with 95% “confidence,” etc. The ratio of # CI’s that contain μ / Total # CI’s $\rightarrow 0.95$, as more and more samples are chosen, i.e., “The probability that a *random* CI contains the population mean μ is equal to 0.95.” In practice however, the common (but technically incorrect) interpretation is that “**the probability that a fixed CI (such as the ones found above) contains μ is 95%.**” In reality, the parameter μ is *constant*; once calculated, a single fixed confidence interval either contains it or not.

For any **significance level** α (and hence **confidence level** $1 - \alpha$), we similarly define the...

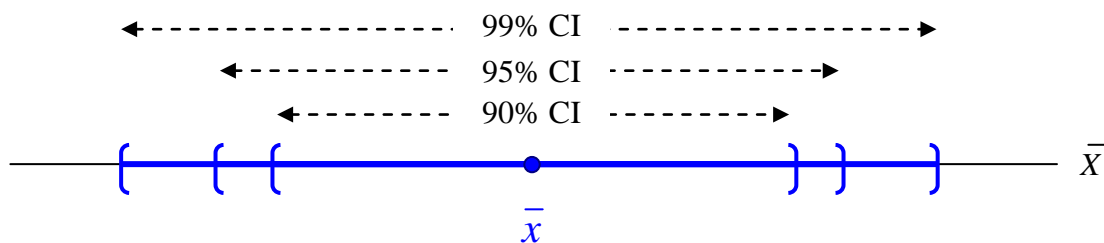
$(1 - \alpha) \times 100\%$ Confidence Interval for μ

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

where $z_{\alpha/2}$ is the **critical value** that divides the area under the **standard normal distribution** $N(0, 1)$ as shown. Recall that for $\alpha = 0.10$, **0.05**, 0.01 (i.e., $1 - \alpha = 0.90$, **0.95**, 0.99), the corresponding critical values are $z_{0.05} = 1.645$, $z_{0.025} = \mathbf{1.960}$, and $z_{0.005} = 2.576$, respectively. The quantity $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ is the *two-sided margin of error*.



Therefore, as the significance level α decreases (i.e., as the confidence level $1 - \alpha$ increases), it follows that the margin of error increases, and thus the corresponding confidence interval widens. Likewise, as the significance level α increases (i.e., as the confidence level $1 - \alpha$ decreases), it follows that the margin of error decreases, and thus the corresponding confidence interval narrows.



Exercise: Why is it not realistic to ask for a 100% confidence interval (i.e., “certainty”)?

Exercise: Calculate the 90% and 99% confidence intervals for Samples 1 and 2 in the preceding example, and compare with the 95% confidence intervals.

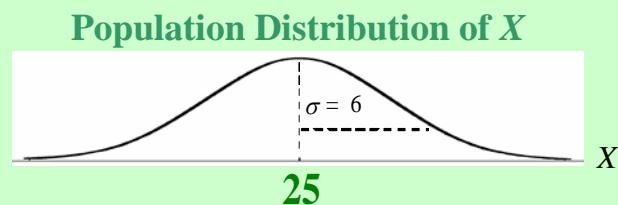
We are now in a position to be able to conduct **Statistical Inference** on the population, via a formal process known as

Objective 2a: Hypothesis Testing ~ “How does this *new* treatment compare with a ‘control’ treatment?” In particular, how can we use a **confidence interval** to decide this?

STANDARD POPULATION = Cancer patients on *standard* drug treatment

Random Variable: X = “Survival time” (months)

Suppose X is known to have **mean = 25** months.



How does this compare with the **mean μ** of the study population?

Technical Notes: Although this is drawn as a bell curve, we don’t really care how the variable X is distributed in *this* population, as long as it is normally distributed in the *study* population of interest, an assumption we will learn how to check later, from the data. Likewise, we don’t really care about the value of the standard deviation σ of *this* population, only of the *study* population. However, in the absence of other information, it is sometimes assumed (not altogether unreasonable) that the two are at least comparable in value. And if this is indeed a *standard* treatment, it has presumably been around for a while and given to many patients, during which time much data has been collected, and thus very accurate parameter estimates have been calculated. Nevertheless, for the vast majority of studies, it is still relatively uncommon that this is the case; in practice, very little if any information is known about any **population** standard deviation σ . In lieu of this value then, σ is usually well-estimated by the **sample** standard deviation s with little change, *if the sample is sufficiently “large,”* but small samples present special problems. These issues will be dealt with later; for now, we will simply assume that the value of σ is known.

Hence, let us consider the situation where, *before any sampling is done*, it is actually the experimenter's intention to see if there is a **statistically significant** difference between the unknown mean survival time μ of the “new treatment” population, and the known mean survival time of 25 months of the “standard treatment” population. (See page 1-1!) That is, the sample data will be used to determine whether or not to reject the formal...

Null Hypothesis $H_0: \mu = 25$

“No significant difference exists.”

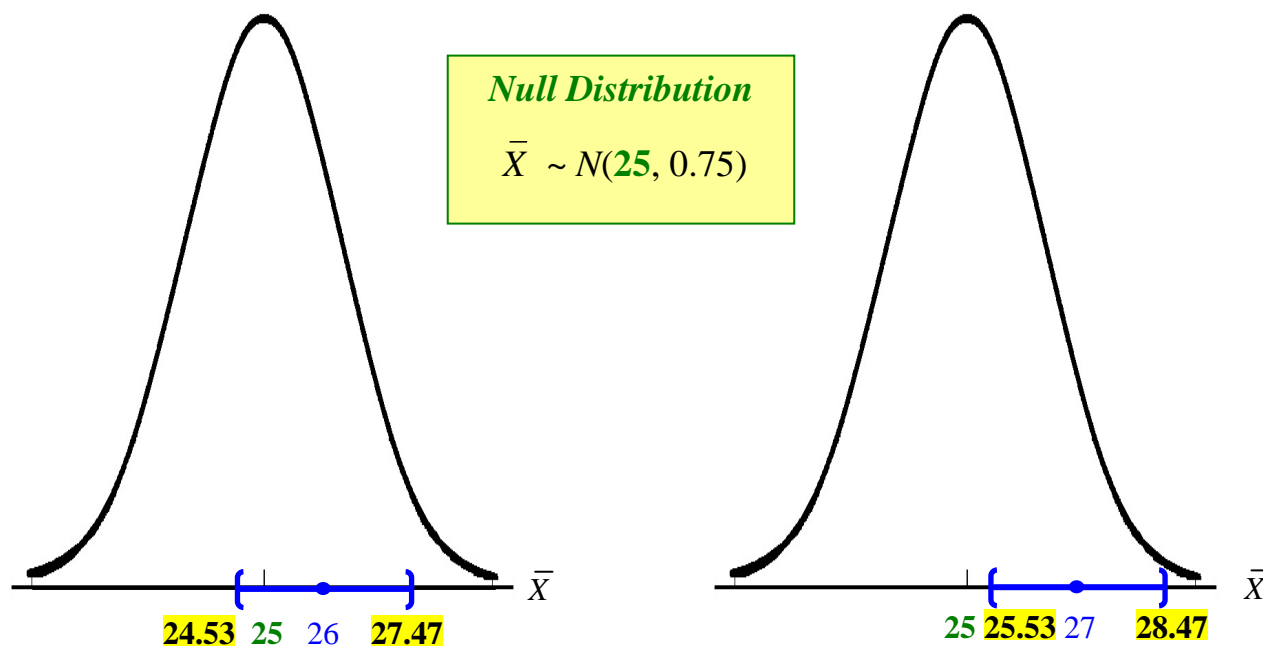
versus the

Alternative Hypothesis $H_A: \mu \neq 25$

Two-sided Alternative

Either $\mu < 25$ or $\mu > 25$

at the $\alpha = 0.05$ significance level (i.e., the 95% confidence level).



Sample 1: 95% CI does contain $\mu = 25$. Therefore, the data support H_0 , and we cannot reject it at the $\alpha = .05$ level. Based on this sample, the new drug does not result in a mean survival time that is **significantly** different from 25 months. *Further study?*

Sample 2: 95% CI does not contain $\mu = 25$. Therefore, the data do not support H_0 , and we can reject it at the $\alpha = .05$ level. Based on this sample, the new drug does result in a mean survival time that is **significantly** different from 25 months. *A genuine treatment effect.*

In general...

Null Hypothesis $H_0: \mu = \mu_0$

versus the

Alternative Hypothesis $H_A: \mu \neq \mu_0$

Two-sided Alternative

Either $\mu < \mu_0$ or $\mu > \mu_0$

Decision Rule: If the $(1 - \alpha) \times 100\%$ confidence interval contains the value μ_0 , then the difference is not statistically significant; “accept” the null hypothesis at the α level of significance. If it does *not* contain the value μ_0 , then the difference is statistically significant; reject the null hypothesis in favor of the alternative at the α significance level.

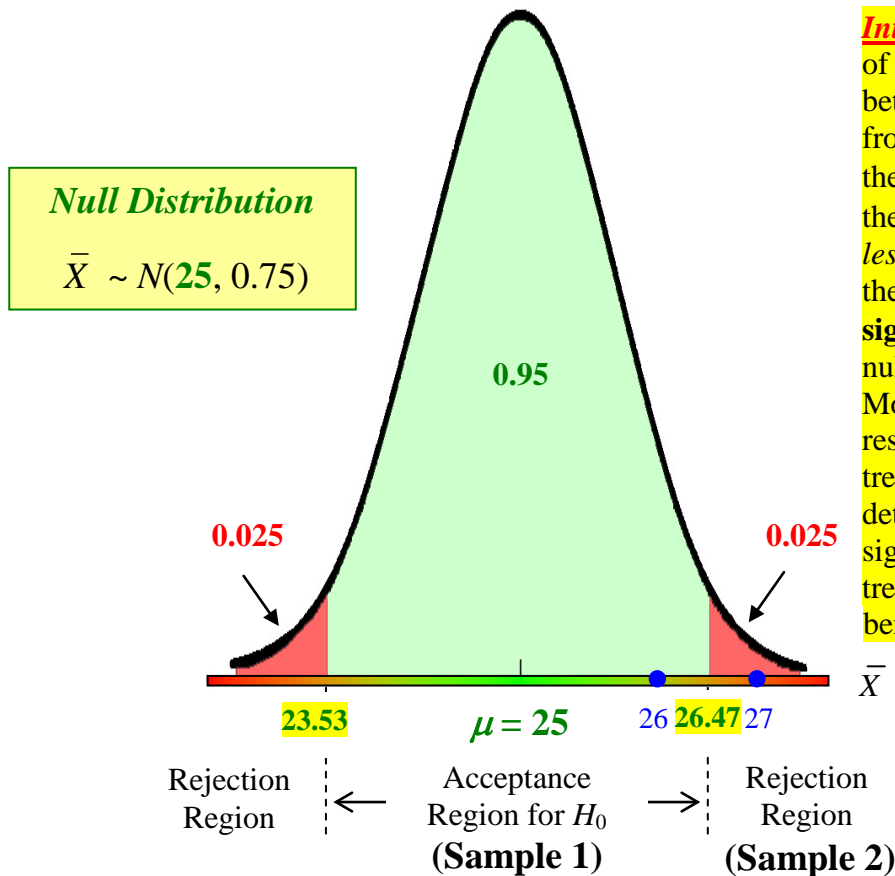
Objective 2b: Calculate which sample mean values \bar{x} will lead to rejecting or not rejecting (i.e., “accepting” or “retaining”) the null hypothesis.

From equation ★ above, and the calculated margin of error = 1.47, we have...

$$P(\mu - 1.47 \leq \bar{X} \leq \mu + 1.47) = 0.95.$$

Now, IF the null hypothesis : $\mu = 25$ is indeed true, then substituting this value gives...

$$P(23.53 \leq \bar{X} \leq 26.47) = 0.95.$$



Interpretation: If the mean survival time \bar{x} of a random sample of $n = 64$ patients is between 23.53 and 26.47, then the difference from 25 is “**not statistically significant**” (at the $\alpha = .05$ significance level), and we **retain** the null hypothesis. However, if \bar{x} is either *less* than 23.53, or *greater* than 26.47, then the difference from 25 will be “**statistically significant**” (at $\alpha = .05$), and we **reject** the null hypothesis in favor of the alternative. More specifically, if the former, then the result is significantly *lower* than the standard treatment average (i.e., new treatment is detrimental!); if the latter, then the result is significantly *higher* than the standard treatment average (i.e., new treatment is beneficial).

In general...

$(1 - \alpha) \times 100\%$ Acceptance Region for $H_0: \mu = \mu_0$

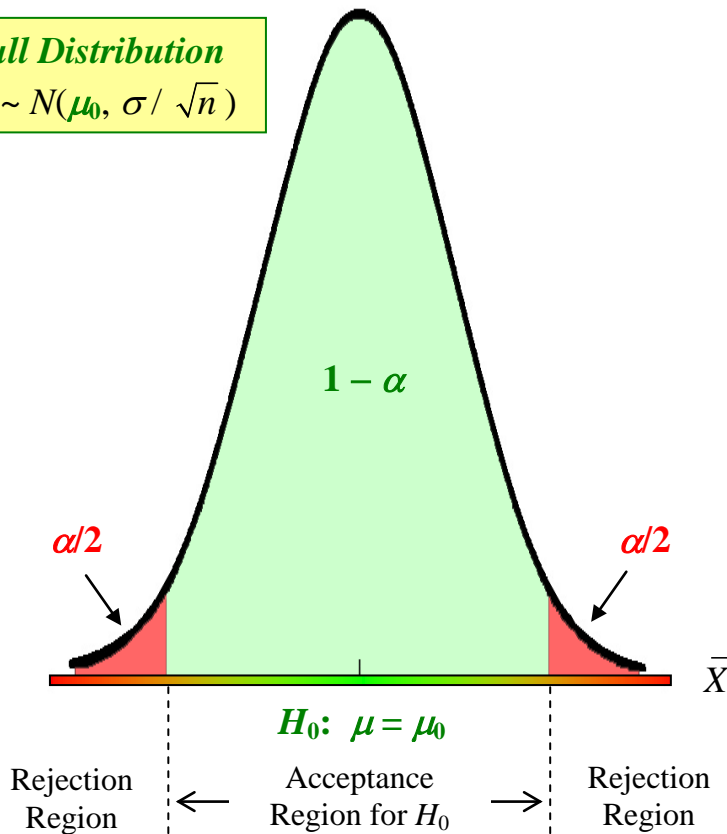
$$\left(\mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Decision Rule: If the $(1 - \alpha) \times 100\%$ acceptance region contains the value \bar{x} , then the difference is not statistically significant; “accept” the null hypothesis at the α significance level. If it does *not* contain the value \bar{x} , then the difference is statistically significant; reject the null hypothesis in favor of the alternative at the α significance level.

Error Rates Associated with Accepting / Rejecting a Null Hypothesis

(vis-à-vis Neyman-Pearson)

Null Distribution
 $\bar{X} \sim N(\mu_0, \sigma / \sqrt{n})$



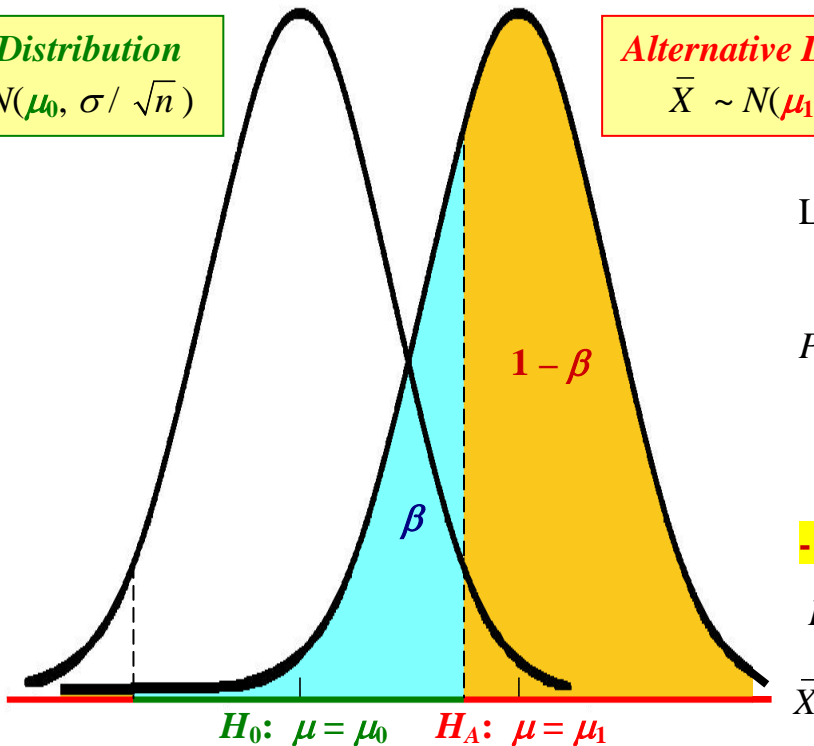
- Confidence Level -

$$P(\text{Accept } H_0 \mid \overbrace{H_0 \text{ true}}^{\mu = \mu_0}) = 1 - \alpha$$

- Significance Level -

$$P(\underbrace{\text{Reject } H_0 \mid H_0 \text{ true}}_{\text{Type I Error}}) = \alpha$$

Null Distribution
 $\bar{X} \sim N(\mu_0, \sigma / \sqrt{n})$



Alternative Distribution
 $\bar{X} \sim N(\mu_1, \sigma / \sqrt{n})$

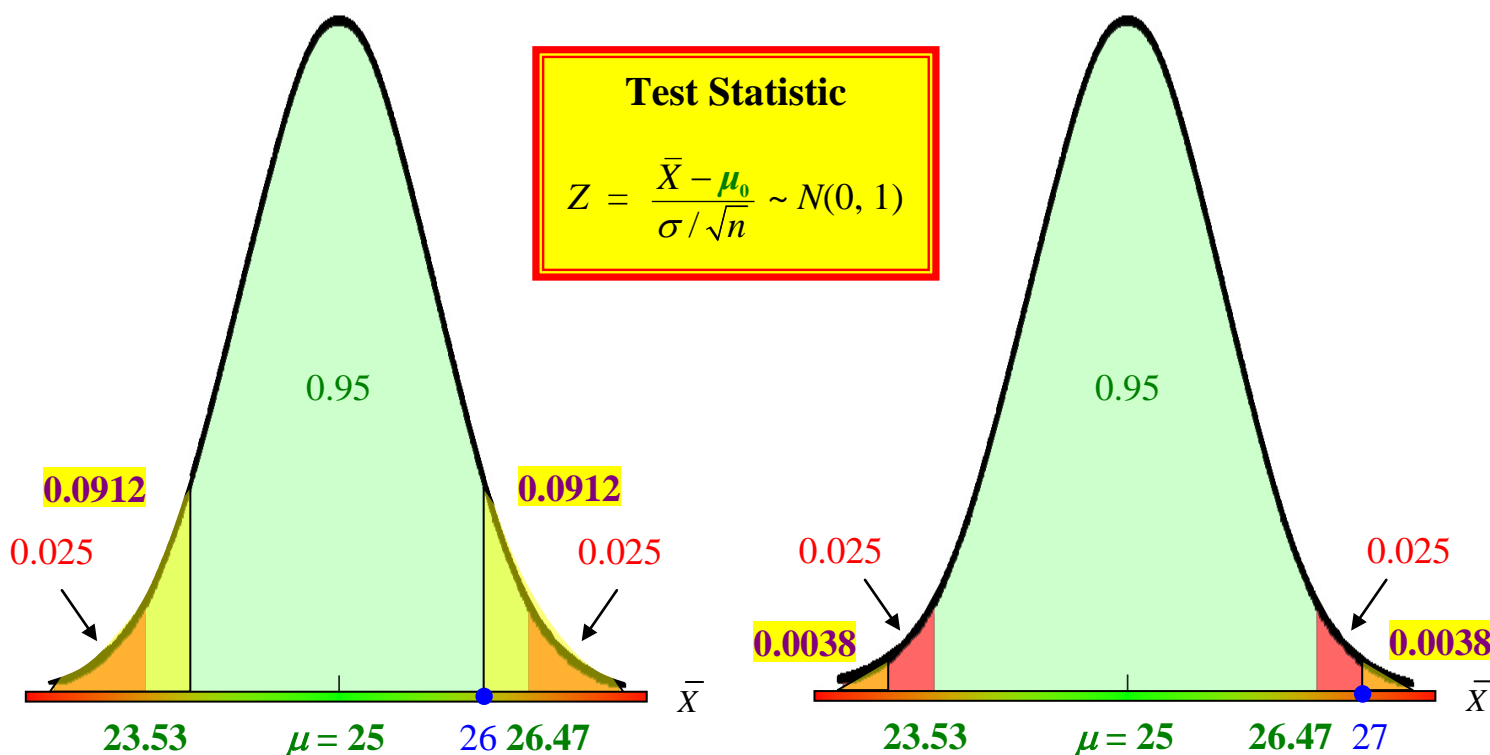
Likewise,

$$P(\underbrace{\text{Accept } H_0 \mid \overbrace{H_0 \text{ false}}^{\mu = \mu_1}}_{\text{Type II Error}}) = \beta$$

- Power -

$$P(\text{Reject } H_0 \mid H_A: \mu = \mu_1) = 1 - \beta$$

Objective 2c: “How *probable* is my experimental result, if the null hypothesis is true?” Consider a sample mean value \bar{x} . Again assuming that the null hypothesis : $\mu = \mu_0$ is indeed true, calculate the **p-value** of the sample = the probability that *any* random sample mean is this far away *or farther*, in the direction of the alternative hypothesis. That is, *how* significant is the decision about H_0 , at level α ?



Sample 1: $p\text{-value} = P(\bar{X} \leq 24 \text{ or } \bar{X} \geq 26)$
 $= P(\bar{X} \leq 24) + P(\bar{X} \geq 26)$
 $= 2 \times P(\bar{X} \geq 26)$
 $= 2 \times P\left(Z \geq \frac{26 - 25}{0.75}\right)$
 $= 2 \times P(Z \geq 1.333)$
 $= 2 \times 0.0912$
 $= \mathbf{0.1824} > 0.05 = \alpha$

Sample 2: $p\text{-value} = P(\bar{X} \leq 23 \text{ or } \bar{X} \geq 27)$
 $= P(\bar{X} \leq 23) + P(\bar{X} \geq 27)$
 $= 2 \times P(\bar{X} \geq 27)$
 $= 2 \times P\left(Z \geq \frac{27 - 25}{0.75}\right)$
 $= 2 \times P(Z \geq 2.667)$
 $= 2 \times 0.0038$
 $= \mathbf{0.0076} < 0.05 = \alpha$

Recall that $Z = 1.96$ is the $\alpha = .05$ cutoff z-score!

Decision Rule: If the p -value of the sample is greater than the significance level α , then the difference is not statistically significant; “accept” the null hypothesis at this level. If the p -value is less than α , then the difference is statistically significant; reject the null hypothesis in favor of the alternative at this level.

Guide to statistical significance of p -values for $\alpha = .05$:

Reject
 H_0

$0 \leq p \leq .001$
extremely strong

$p \approx .005$
strong

$p \approx .01$
moderate

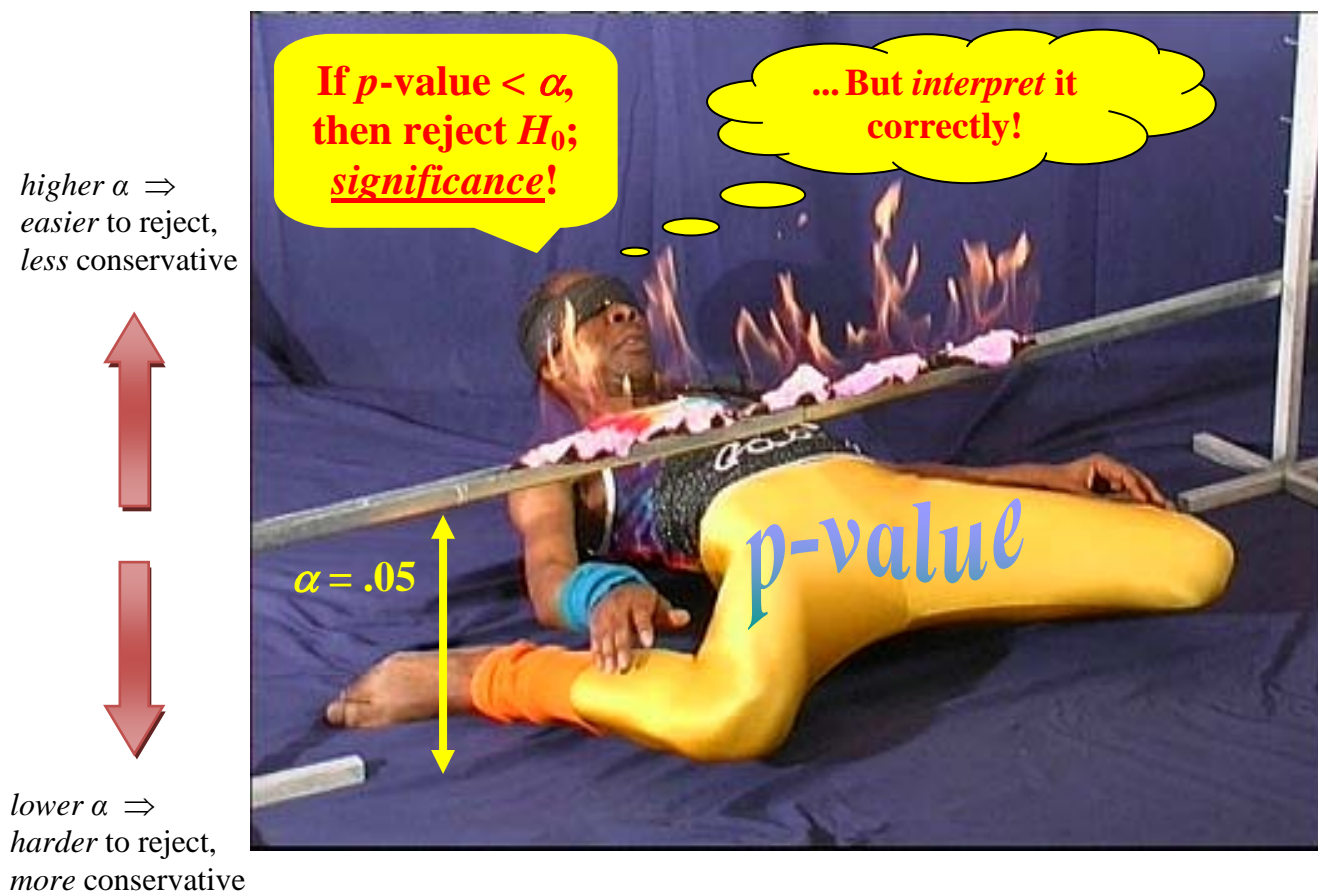
$p \approx .05$
borderline

$.10 \leq p \leq 1$
not significant

Accept
 H_0

Summary of findings: Even though the data from both samples suggest a generally longer “mean survival time” among the “new treatment” population over the “standard treatment” population, the formal conclusions and interpretations are different. Based on **Sample 1** patients ($\bar{x} = 26$), the difference between the mean survival time μ of the study population, and the mean survival time of 25 months of the standard population, is **not statistically significant**, and may in fact simply be due to random chance. Based on **Sample 2** patients ($\bar{x} = 27$) however, the difference between the mean age μ of the study population, and the mean age of 25 months of the standard population, is indeed **statistically significant**, on the *longer* side. Here, the *increased* survival times serve as empirical evidence of a genuine, beneficial “**treatment effect**” of the new drug.

Comment: For the sake of argument, suppose that a third sample of patients is selected, and to the experimenter’s surprise, the sample mean survival time is calculated to be only $\bar{x} = 23$ months. Note that the p -value of this sample is *the same as* Sample 2, with $\bar{x} = 27$ months, namely, $0.0076 < 0.05 = \alpha$. Therefore, as far as **inference** is concerned, the formal conclusion is the same, namely, reject H_0 : $\mu = 25$ months. However, the practical **interpretation** is very different! While we do have statistical significance as before, these patients survived considerably *shorter* than the standard average, i.e., the treatment had an unexpected effect of *decreasing* survival times, rather than increasing them. (This kind of unanticipated result is more common than you might think, especially with investigational drugs, which is one reason for formal hypothesis testing, before drawing a conclusion.)



Modification: Consider now the (unlikely?) situation where the experimenter knows that the new drug will not result in a “mean survival time” μ that is significantly less than 25 months, and would specifically like to determine if there is a **statistically significant increase**. That is, he/she formulates the following **one-sided** null hypothesis to be rejected, and complementary alternative:

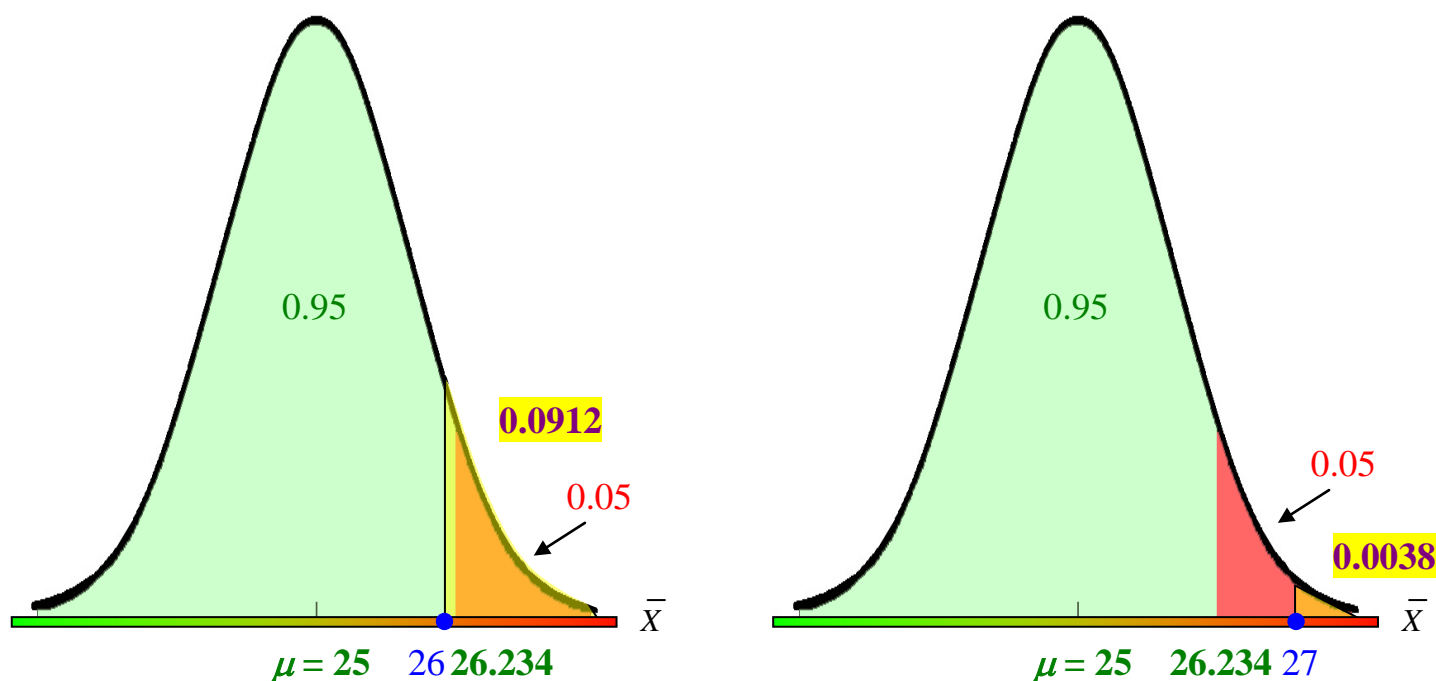
Null Hypothesis $H_0: \mu \leq 25$

versus the

Alternative Hypothesis $H_A: \mu > 25$ Right-tailed Alternative

at the $\alpha = 0.05$ significance level (i.e., the 95% confidence level).

In this case, the **acceptance region** for H_0 consists of sample mean values \bar{x} that are *less* than the null-value of $\mu_0 = 25$, plus the *one-sided margin of error* $= z_{\alpha} \frac{\sigma}{\sqrt{n}} = z_{.05} \frac{6}{\sqrt{64}} = (1.645)(0.75) = 1.234$, hence **26.234**. **Note that α replaces $\alpha/2$ here!**



Sample 1: $p\text{-value} = P(\bar{X} \geq 26)$

$$= P(Z \geq 1.333)$$

$$= \mathbf{0.0912} > 0.05 = \alpha$$

(accept)

Sample 2: $p\text{-value} = P(\bar{X} \geq 27)$

$$= P(Z \geq 2.667)$$

$$= \mathbf{0.0038} < 0.05 = \alpha$$

(fairly strong rejection)

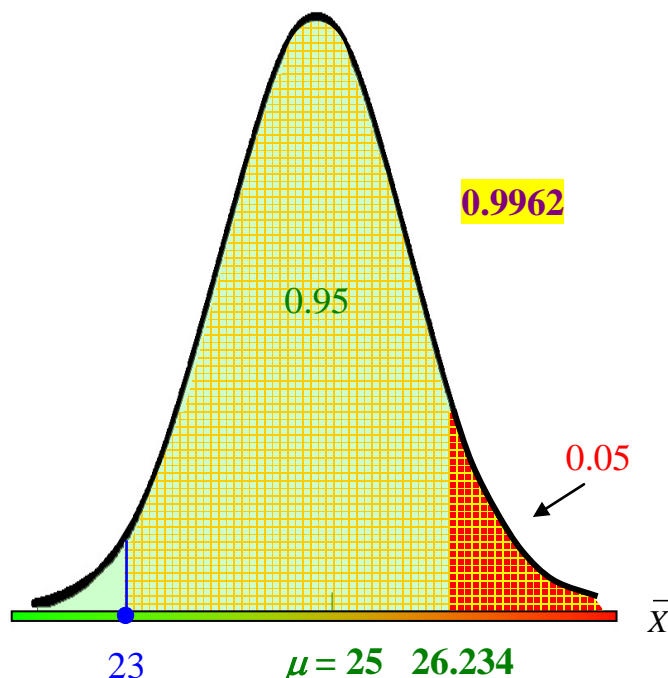
Here, $Z = 1.645$ is the $\alpha = .05$ cutoff z-score! Why?

Note that these one-sided p -values are exactly *half* of their corresponding two-sided p -values found above, potentially making the null hypothesis easier to reject. However, there are subtleties that arise in one-sided tests that do not arise in two-sided tests...

Consider again the third sample of patients, whose sample mean is unexpectedly calculated to be only $\bar{x} = 23$ months. Unlike the previous two samples, this evidence is in *strong agreement* with the null hypothesis $H_0: \mu \leq 25$ that the “mean survival time” is 25 months or less. This is confirmed by the p -value of the sample, whose definition (recall above) is “the probability that *any* random sample mean is this far away *or farther*, in the direction of the alternative hypothesis” which, in this case, is the *right-sided* $H_A: \mu > 25$. Hence,

$$p\text{-value} = P(\bar{X} \geq 23) = P(Z \geq -2.667) = 1 - 0.0038 = 0.9962 \gg 0.05 = \alpha$$

which, as just observed informally, indicates a strong “acceptance” of the null hypothesis.



Exercise: What is the one-sided p -value if the sample mean $\bar{x} = 24$ mos? Conclusions?

A word of caution: One-sided tests are less **conservative** than two-sided tests, and should be used sparingly, especially when it is *a priori* unknown if the mean response μ is likely to be significantly larger or smaller than the null-value μ_0 , e.g., testing the effect of a new drug. More appropriate to use when it can be clearly assumed from the circumstances that the conclusion would only be of *practical* significance if μ is either higher or lower (but not both) than some tolerance or threshold level μ_0 , e.g., toxicity testing, where only higher levels are of concern.

SUMMARY: To test *any* null hypothesis for one mean μ , via the p -value of a sample...

- Step I: Draw a **picture** of a bell curve, centered at the “null value” μ_0 .
- Step II: Calculate your sample mean \bar{x} , and **plot it** on the horizontal \bar{X} axis.
- Step III: From \bar{x} , find the area(s) **in the direction(s) of H_A ($<$, $>$, or both tails)**, by first transforming \bar{x} to a **z -score**, and using the z -table. This is your **p -value**. **SEE NEXT PAGE!**
- Step IV: Compare p with the significance level α . If $<$, **reject** H_0 . If $>$, **retain** H_0 .
- Step V: **Interpret** your conclusion in the context of the given situation!

P-VALUES MADE EASY

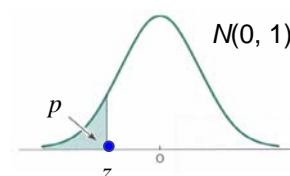
Def: Suppose a null hypothesis H_0 about a population mean μ is to be tested, at a significance level α ($= .05$, usually), using a known sample mean \bar{x} from an experiment. The **p-value** of the sample is the probability that a *general* random sample yields a mean \bar{X} that differs from the hypothesized “null value” μ_0 , by an amount which is as large as – or larger than – the difference between our known \bar{x} value and μ_0 .

Thus, a small p -value (i.e., $< \alpha$) indicates that our sample provides evidence against the null hypothesis, and we may *reject* it; the smaller the p -value, the stronger the rejection, and the more “statistically significant” the finding. A p -value $> \alpha$ indicates that our sample does not provide evidence against the null hypothesis, and so we may not reject it. Moreover, a large p -value (i.e., ≈ 1) indicates empirical evidence in support of the null hypothesis, and we may retain, or even “*accept*” it. Follow these simple steps:

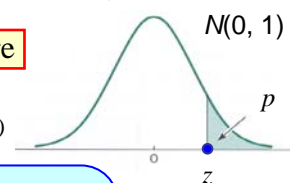
STEP 1. From your *sample mean* \bar{x} , calculate the standardized **z-score** $= \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$. “standard error”

STEP 2. What form is your *alternative hypothesis*?

$H_A: \mu < \mu_0$ (1-sided, *left*)..... **p-value** = tabulated entry corresponding to z -score
= left shaded area, whether $z < 0$ or $z > 0$
(illustrated)



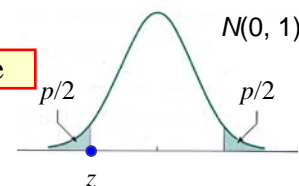
$H_A: \mu > \mu_0$ (1-sided, *right*)..... **p-value** = 1 – tabulated entry corresponding to z -score
= right shaded area, whether $z < 0$ or $z > 0$
(illustrated)



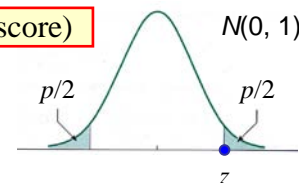
Example: Toxic levels of arsenic in drinking water? Test $H_0: \mu < 10$ ppb (safe) vs. $H_A: \mu \geq 10$ ppb (unsafe), at $\alpha = .05$. Assume $N(\mu, \sigma)$, with $\sigma = 1.6$ ppb. A sample of $n = 64$ readings that average to $\bar{x} = 10.1$ ppb would have a z -score $= 0.1 / 0.2 = 0.5$, which corresponds to a **p-value** $= 1 - 0.69146 = 0.30854 > .05$, hence not significant; toxicity has *not* been formally shown. (Unsafe levels are $\bar{x} \geq 10.33$ ppb. Why?)

$H_A: \mu \neq \mu_0$ (2-sided)

- If z -score is *negative*..... **p-value** = $2 \times$ tabulated entry corresponding to z -score
= $2 \times$ left-tailed shaded area



- If z -score is *positive*..... **p-value** = $2 \times (1 - \text{tabulated entry corresponding to } z\text{-score})$
= $2 \times$ right-tailed shaded area



STEP 3.

- If the p -value is *less* than α ($= .05$, usually), then **REJECT NULL HYPOTHESIS – EXPERIMENTAL RESULT IS STATISTICALLY SIGNIFICANT AT THIS LEVEL!**
- If the p -value is *greater* than α ($= .05$, usually), then **RETAIN NULL HYPOTHESIS – EXPERIMENTAL RESULT IS NOT STATISTICALLY SIGNIFICANT AT THIS LEVEL!**

STEP 4. IMPORTANT - Interpret results in context. (Note: For many, this is the hardest step of all!)

P-VALUES MADE SUPER EASY

STATBOT 301, MODEL Z

SUBJECT: BASIC CALCULATION OF P-VALUES FOR Z-TEST

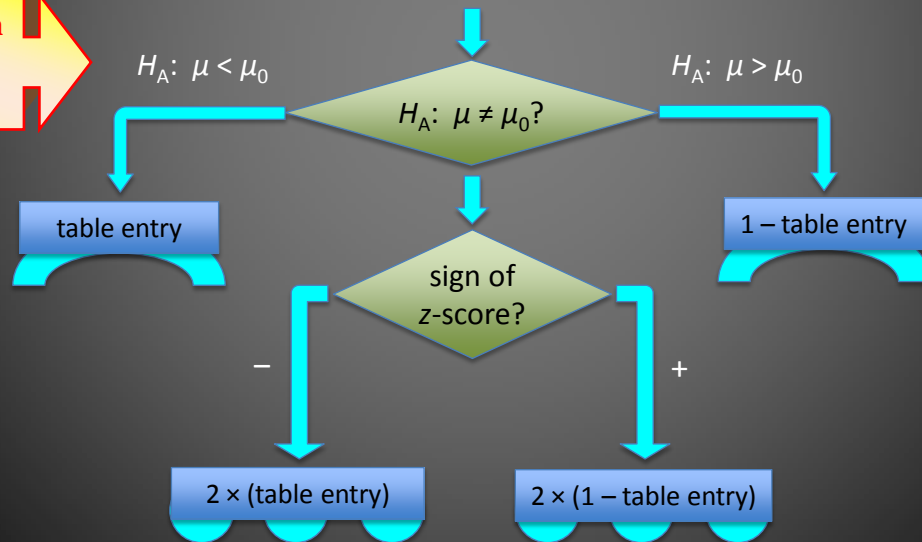
CALCULATE... from H_0

Test Statistic

$$\text{"z-score"} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Remember that the Z-table corresponds to the "cumulative" area to the left of any z-score.

Check the direction of the alternative hypothesis!



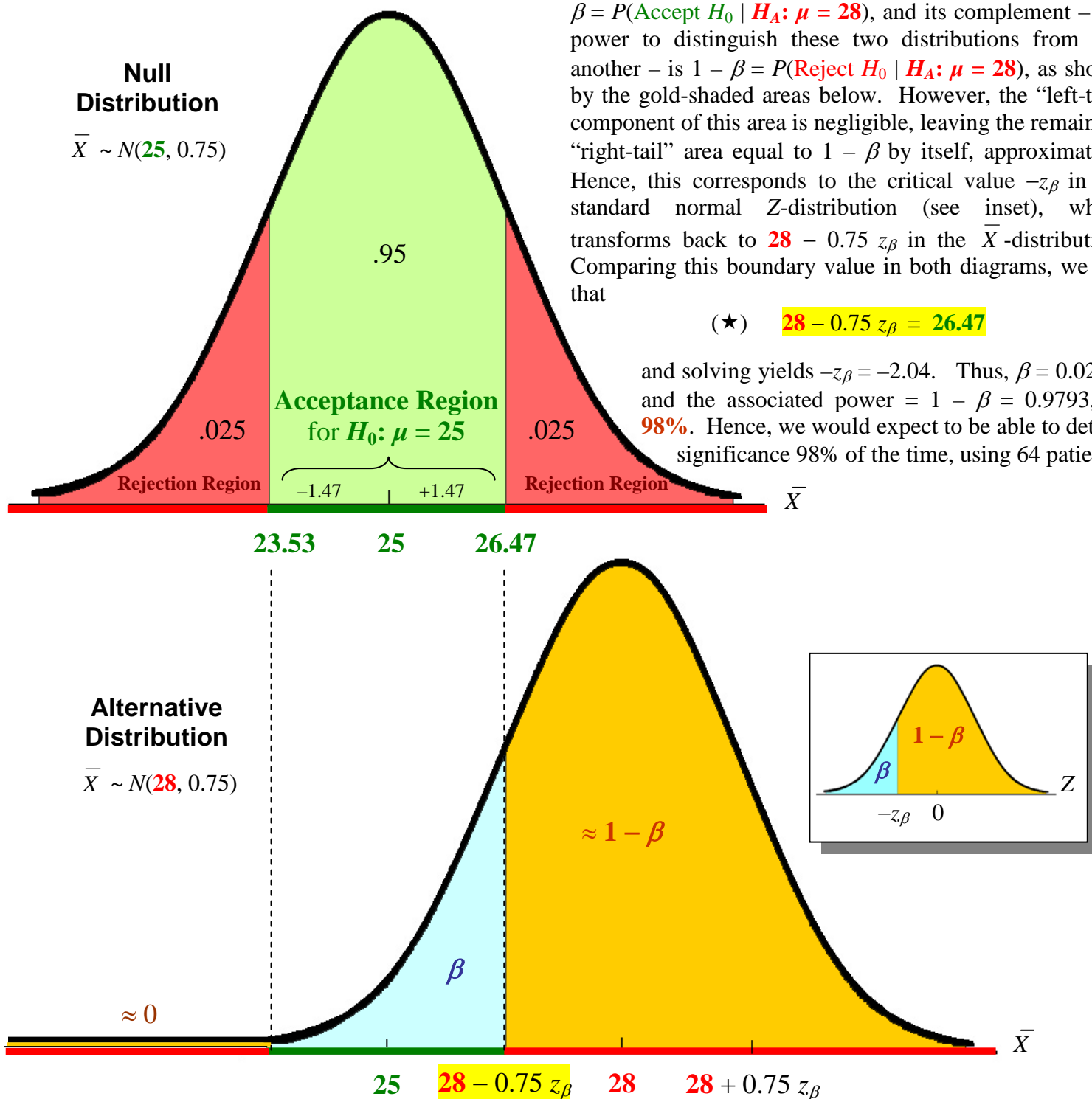
Power and Sample Size Calculations

Recall: $X = \text{survival time (mos)} \sim N(\mu, \sigma)$, with $\sigma = 6$ (given). Testing null hypothesis $H_0: \mu = 25$ (versus the 2-sided alternative $H_A: \mu \neq 25$), at the $\alpha = .05$ significance level. Also recall that, by definition, **power** $= 1 - \beta = P(\text{Reject } H_0 \mid H_0 \text{ is false, i.e., } \mu \neq 25)$. Indeed, suppose that the mean survival time of “new treatment” patients is actually suspected to be $H_A: \mu = 28$ mos. In this case, what is the resulting power to distinguish the difference and reject H_0 , using a sample of $n = 64$ patients (as in the previous examples)?

These diagrams compare the null distribution for $\mu = 25$, with the alternative distribution corresponding to $\mu = 28$ in the rejection region of the null hypothesis. By definition, $\beta = P(\text{Accept } H_0 \mid H_A: \mu = 28)$, and its complement – the power to distinguish these two distributions from one another – is $1 - \beta = P(\text{Reject } H_0 \mid H_A: \mu = 28)$, as shown by the gold-shaded areas below. However, the “left-tail” component of this area is negligible, leaving the remaining “right-tail” area equal to $1 - \beta$ by itself, approximately. Hence, this corresponds to the critical value $-z_\beta$ in the standard normal Z-distribution (see inset), which transforms back to $28 - 0.75 z_\beta$ in the \bar{X} -distribution. Comparing this boundary value in both diagrams, we see that

$$(\star) \quad 28 - 0.75 z_\beta = 26.47$$

and solving yields $-z_\beta = -2.04$. Thus, $\beta = 0.0207$, and the associated power $= 1 - \beta = 0.9793$, or **98%**. Hence, we would expect to be able to detect significance 98% of the time, using 64 patients.



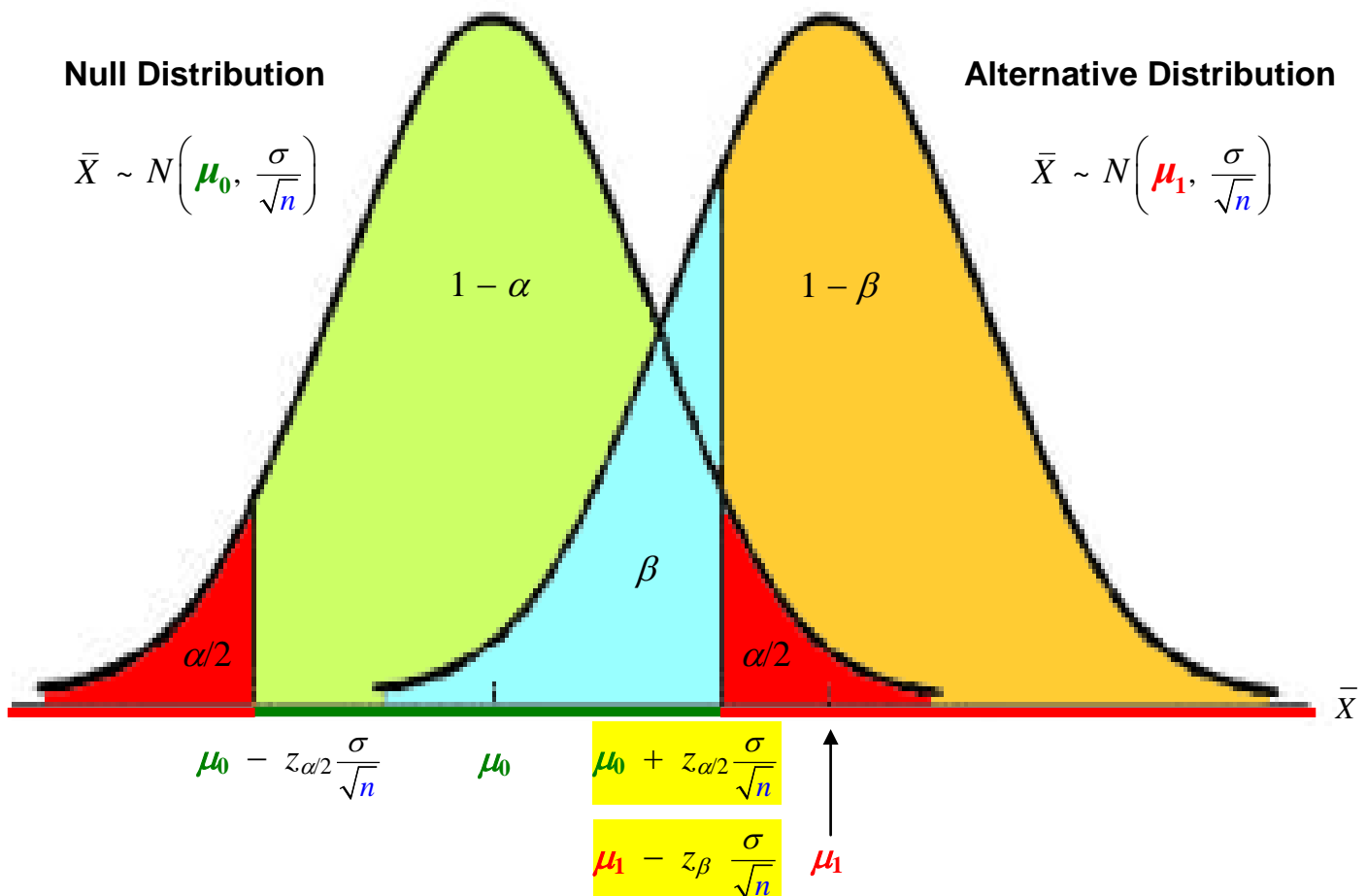
General Formulation:

Procurement of drug samples for testing purposes, or patient recruitment for clinical trials, can be extremely time-consuming and expensive. How to determine the minimum sample size n required to reject the null hypothesis $H_0: \mu = \mu_0$, in favor of an alternative value $H_A: \mu = \mu_1$, with a desired power $1 - \beta$, at a specified significance level α ? (And conversely, how to determine the power $1 - \beta$ for a given sample size n , as above?)

	H_0 true	H_0 false
Reject H_0	✗ Type I error, probability = α (significance level)	✓ probability = $1 - \beta$ (power)
Accept H_0	✓ probability = $1 - \alpha$ (confidence level)	✗ Type II error, probability = β (1 - power)

That is, **confidence level** = $1 - \alpha = P(\text{Accept } H_0: \mu = \mu_0 \mid H_0 \text{ is true})$,

and **power** = $1 - \beta = P(\text{Reject } H_0: \mu = \mu_0 \mid H_A: \mu = \mu_1)$.



Hence (compare with (★) above),

$$\mu_1 - z_\beta \frac{\sigma}{\sqrt{n}} = \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

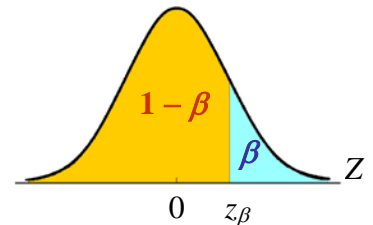
Solving for n yields the following.

In order to be able to detect a statistically significant difference (at level α) between the null population distribution having mean μ_0 , and an alternative population distribution having mean μ_1 , with a **power** of $1 - \beta$, we require a *minimum* sample size of

$$n = \left(\frac{z_{\alpha/2} + z_\beta}{\Delta} \right)^2,$$

where $\Delta = \frac{|\mu_1 - \mu_0|}{\sigma}$ is the “scaled difference” between μ_0 and μ_1 .

Note: Remember that, as we defined it, z_β is always ≥ 0 , and has β area to its right.

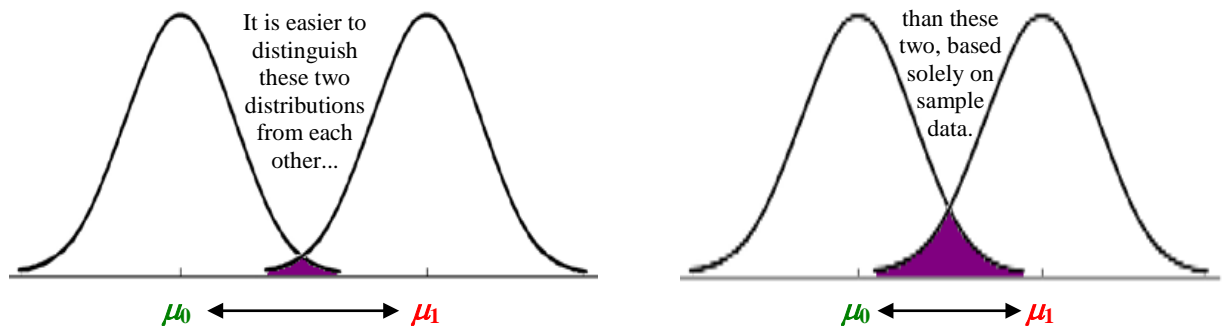


Comments:

- This formula corresponds to a *two-sided* hypothesis test. For a *one-sided* test, simply replace $\alpha/2$ by α . Recall that if $\alpha = .05$, then $z_{.025} = 1.960$ and $z_{.05} = 1.645$.
- If σ is not known, then it can be replaced above by s , the sample standard deviation, provided the resulting sample size turns out to be $n \geq 30$, to be consistent with CLT. However, if the result is $n < 30$, then add 2 to compensate. [Modified from: Lachin, J. M. (1981), Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2(2), 93-113.]

What affects sample size, and how? With all other values being equal...

- As power $1 - \beta$ increases, n increases; as $1 - \beta$ decreases, n decreases.
- As the difference Δ decreases, n increases; as Δ increases, n decreases.



Exercise: Also show that n increases...

- as σ increases, [Hint: It may be useful to draw a picture, similar to above.]
- as α decreases. [Hint: It may be useful to recall that α is the Type I Error rate, or equivalently, that $1 - \alpha$ is the confidence level.]

Examples: Recall that in our study, $\mu_0 = 25$ months, $\sigma = 6$ months.

Suppose we wish to detect a statistically significant difference (at level $\alpha = .05 \Rightarrow z_{.025} = 1.960$) between this null distribution, and an alternative distribution having...

■ $\mu_1 = 28$ months, with **90% power** ($1 - \beta = .90 \Rightarrow \beta = .10 \Rightarrow z_{.10} = 1.282$). Then the scaled difference $\Delta = \frac{|28 - 25|}{6} = 0.5$, and

$$n = \left(\frac{1.960 + 1.282}{0.5} \right)^2 = 42.04, \quad \text{so} \quad n \geq 43 \text{ patients.}$$

■ $\mu_1 = 28$ months, with **95% power** ($1 - \beta = .95 \Rightarrow \beta = .05 \Rightarrow z_{.05} = 1.645$). Then,

$$n = \left(\frac{1.960 + 1.645}{0.5} \right)^2 = 51.98, \quad \text{so} \quad n \geq 52 \text{ patients.}$$

■ $\mu_1 = 27$ months, with **95% power** (so again, $z_{.05} = 1.645$). Then $\Delta = \frac{|27 - 25|}{6} = 0.333$,

$$n = \left(\frac{1.960 + 1.645}{0.333} \right)^2 = 116.96, \quad \text{so} \quad n \geq 117 \text{ patients.}$$

Table of Sample Sizes* for Two-Sided Tests ($\alpha = .05$)

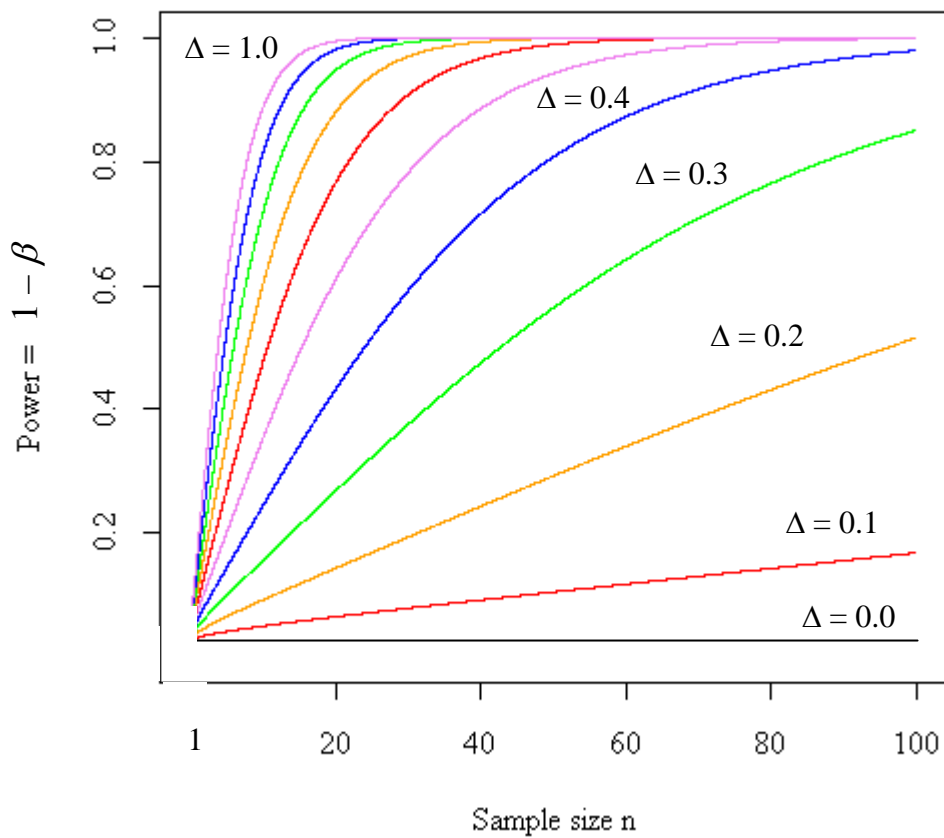
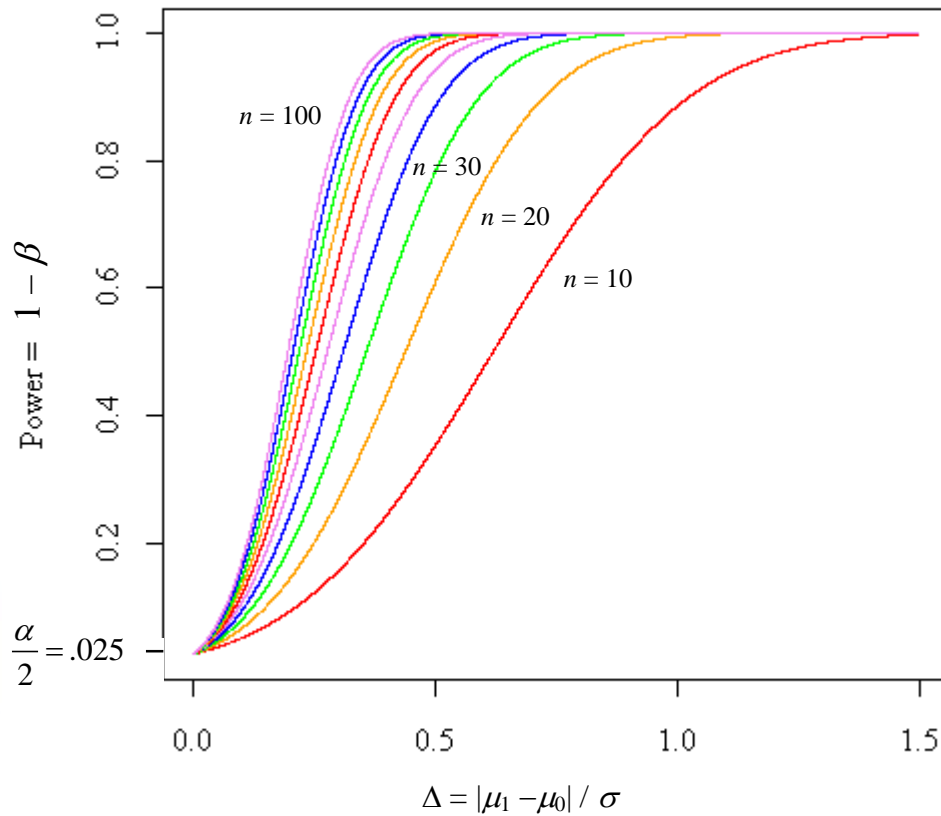
Δ	Power				
	80%	85%	90%	95%	99%
0.1	785	898	1051	1300	1838
0.125	503	575	673	832	1176
0.15	349	400	467	578	817
0.175	257	294	344	425	600
0.2	197	225	263	325	460
0.25	126	144	169	208	294
0.3	88	100	117	145	205
0.35	65	74	86	107	150
0.4	50	57	66	82	115
0.45	39	45	52	65	91
0.5	32	36	43	52	74
0.6	24	27	30	37	52
0.7	19	21	24	29	38
0.8	15	17	19	23	31
0.9	12	14	15	19	25
1.0	10	11	13	15	21

* Shaded cells indicate that 2 was added to compensate for small n .

Power Curves – A visual way to relate power and sample size.

Question:

Why is power not equal to 0 if $\Delta = 0$?



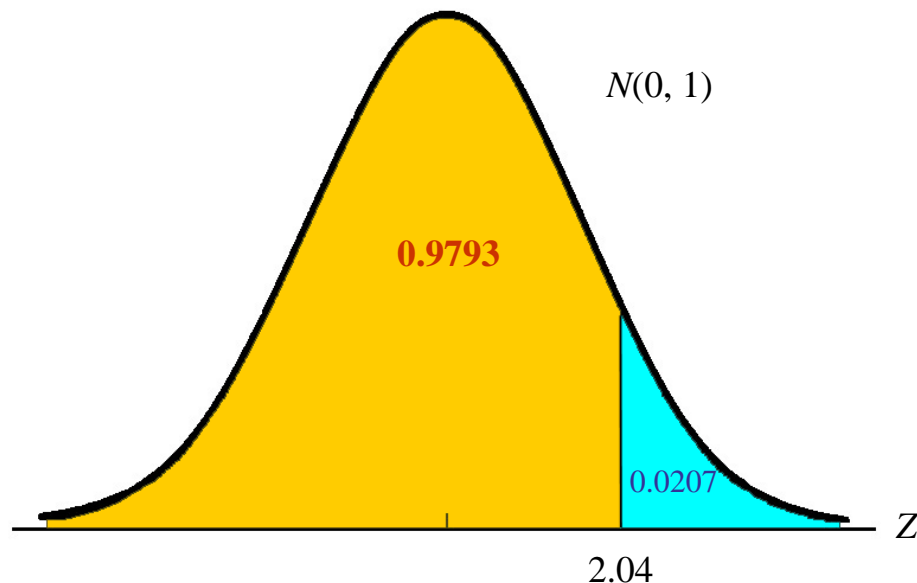
Comments:

- Due to time and/or budget constraints for example, a study may end before optimal sample size is reached. Given the current value of n , the corresponding power can then be determined by the graph above, or computed exactly via the following formula.

$$\text{Power} = 1 - \beta = P(Z \leq \underbrace{-z_{\alpha/2} + \Delta \sqrt{n}}_{\text{z-score}})$$

The **z-score** can be +, -, or 0.

Example: As in the original study, let $\alpha = .05$, $\Delta = \frac{|28 - 25|}{6} = 0.5$, and $n = 64$. Then the **z-score** $= -1.96 + 0.5\sqrt{64} = 2.04$, so power $= 1 - \beta = P(Z \leq 2.04) = \mathbf{0.9793}$, or 98% . The probability of committing a Type 2 error $= \beta = 0.0207$, or 2%. **See page 6.1-15.**



Exercise: How much power exists if the sample size is $n = 25$? 16? 9? 4? 1?

- Generally, a minimum of 80% power is acceptable for reporting purposes.
- Note: Larger sample size \Rightarrow longer study time \Rightarrow longer wait for results. In clinical trials and other medical studies, formal protocols exist for early study termination.
- Also, to achieve a target sample size, practical issues must be considered (e.g., parking, meals, bed space,...). Moreover, may have to recruit many more individuals due to eventual **censoring** (e.g., move-aways, noncompliance,...) or **death**. \$\$\$\$\$\$ issues...
- Research proposals must have power and sample size calculations in their “methods” section, in order to receive institutional approval, support, and eventual journal publication.

Small Samples: Student's t -distribution

Recall that, vis-à-vis the **Central Limit Theorem**: $X \sim N(\mu, \sigma) \Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, for any n .

Test statistic...

- σ known: $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$
- σ unknown, $n \geq 30$: $Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1)$ approximately
- σ unknown, $n < 30$: $T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$ ← **Note:** Can use for $n \geq 30$ as well.

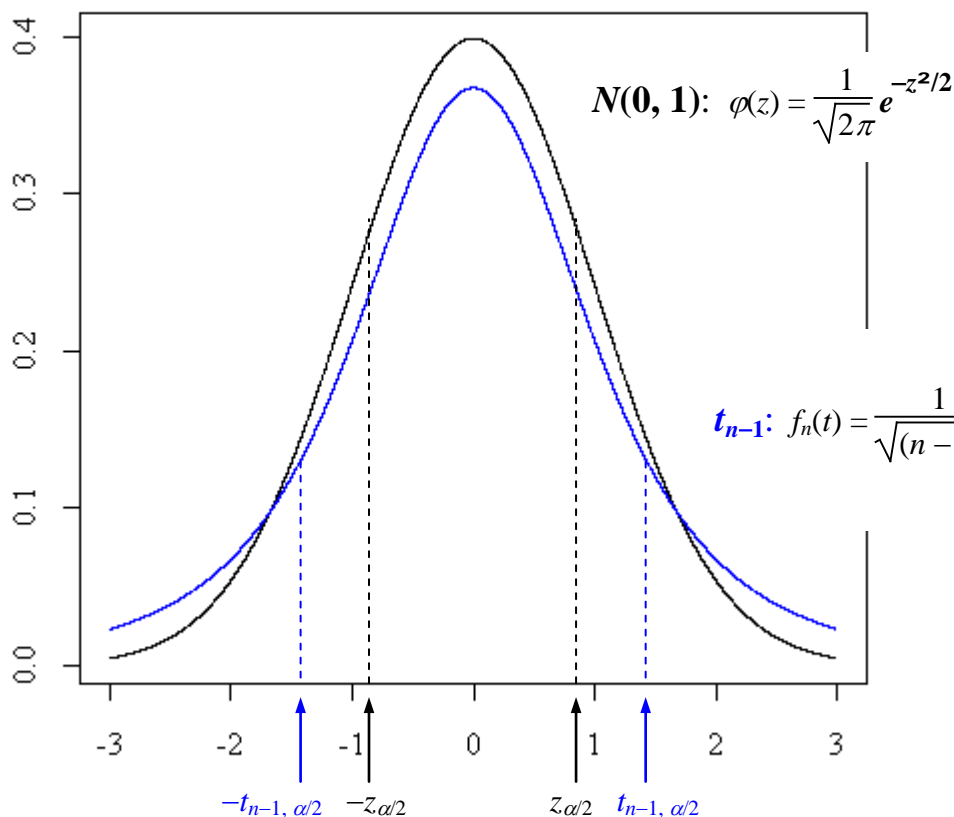
Recall:

$$\text{s.e.} = \sigma / \sqrt{n}$$

$$\widehat{\text{s.e.}} = s / \sqrt{n}$$

Student's t -distribution, with $\nu = n - 1$ **degrees of freedom** $df = 1, 2, 3, \dots$

(Due to William S. Gossett (1876 - 1937), Guinness Brewery, Ireland, anonymously publishing under the pseudonym “Student” in 1908.)



$df = 1$ is also known as the **Cauchy distribution**.

As $df \rightarrow \infty$, it follows that $T \sim t_{df} \rightarrow Z \sim N(0, 1)$.

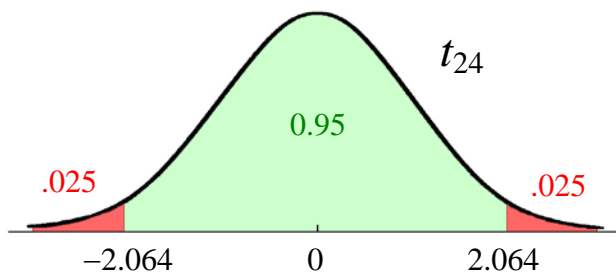
Example: Again recall that in our study, the variable X = “survival time” was assumed to be normally distributed among cancer patients, with $\sigma = 6$ months. The null hypothesis $H_0: \mu = 25$ months was tested with a random sample of $n = 64$ patients; a sample mean of $\bar{x} = 27.0$ months was shown to be statistically significant ($p = .0076$), i.e., sufficient evidence to reject the null hypothesis, suggesting a genuine difference, at the $\alpha = .05$ level.

Now suppose that σ is unknown and, like μ , must also be estimated from sample data. Further suppose that the sample size is small, say $n = 25$ patients, with which to test the same null hypothesis $H_0: \mu = 25$, versus the two-sided alternative $H_A: \mu \neq 25$, at the $\alpha = .05$ significance level. Imagine that a sample mean $\bar{x} = 27.4$ months, and a sample standard deviation $s = 6.25$ months, are obtained. The greater mean survival time appears promising. However...

$$\widehat{\text{s.e.}} = \frac{s}{\sqrt{n}} = \frac{6.25 \text{ mos}}{\sqrt{25}} = 1.25 \text{ months}$$

($> \text{s.e.} = 0.75 \text{ months}$)

$$\text{critical value} = t_{24, .025} = 2.064$$

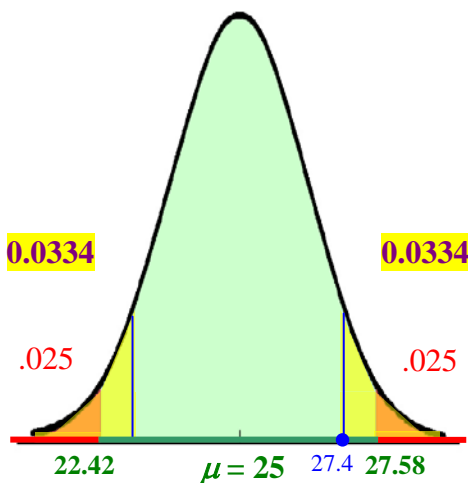


Therefore,

$$\begin{aligned} \text{Margin of Error} &= (2.064)(1.25 \text{ mos}) \\ &= \mathbf{2.58 \text{ months}} \\ &(> 1.47 \text{ months, previously}) \end{aligned}$$

So...

- **95% Confidence Interval for μ** = $(27.4 - 2.58, 27.4 + 2.58) = \mathbf{(24.82, 29.98) \text{ months}}$, which does contain the null value $\mu = 25 \Rightarrow$ Accept H_0 ... No significance shown!
- **95% Acceptance Region for H_0** = $(25 - 2.58, 25 + 2.58) = \mathbf{(22.42, 27.58) \text{ months}}$, which does contain the sample mean $\bar{x} = 27.4 \Rightarrow$ Accept H_0 ... No significance shown!



- **p-value** = $2 P(\bar{X} \geq 27.4) = 2 P\left(T_{24} \geq \frac{27.4 - 25}{1.25}\right)$
 $= 2 P(T_{24} \geq 1.92) = 2(0.0334) = \mathbf{0.0668}$, which is greater than $\alpha = .05 \Rightarrow$ Accept H_0 ... No significance shown!

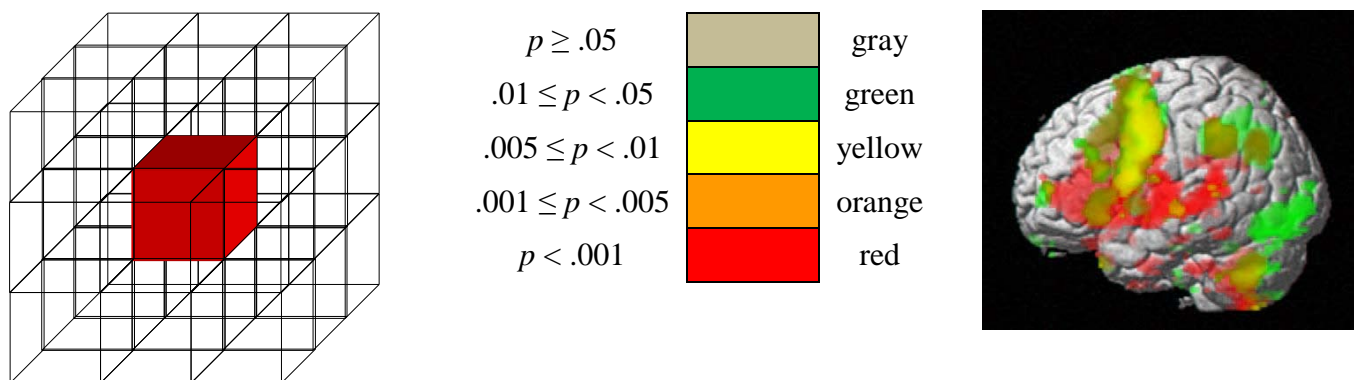
Why? The inability to reject is a typical consequence of small sample size, thus low power!

Also see [Appendix > Statistical Inference > Mean, One Sample](#) for more info and many more examples on this material.

Example: A very simplified explanation of how fMRI works

Functional Magnetic Resonance Imaging (fMRI) is one technique of visually mapping areas of the human cerebral cortex in real time. First, a three-dimensional computer-generated image of the brain is divided into cube-shaped **voxels** (i.e., “volume elements” – analogous to square “picture elements,” or **pixels**, in a two-dimensional image), about 2-4 mm on a side, each voxel containing thousands of neurons. While the patient is asked to concentrate on a specific mental task, increased cerebral blood flow releases oxygen to activated neurons at a greater rate than to inactive ones (the so-called “hemodynamic response”), and the resulting magnetic resonance signal can be detected. In one version, each voxel signal is compared with the mean of its neighboring voxels; if there is a statistically significant difference in the measurements, then the original voxel is assigned one of several colors, depending on the intensity of the signal (e.g., as determined by the p -value); see figures.

Suppose the variable X = “Cerebral Blood Flow (CBF)” typically follows a normal distribution with mean $\mu = 0.5$ ml/g/min at baseline. Further, suppose that the $n = 6$ neighbors surrounding a particular voxel (i.e., front and back, left and right, top and bottom) yields a sample mean of $\bar{x} = 0.767$ ml/g/min, and sample standard deviation of $s = 0.082$ ml/g/min. Calculate the *two-sided* p -value of this sample (using baseline as the null hypothesis for simplicity), and determine what color should be assigned to the central voxel, using the scale shown.



Solution: X = “Cerebral Blood Flow (CBF)” is normally distributed, $H_0: \mu = 0.5$ ml/g/min
 $n = 6$ $\bar{x} = 0.767$ ml/g/min $s = 0.082$ ml/g/min

As the population standard deviation σ is unknown, and the sample size n is small, the t -test on $df = 6 - 1 = 5$ degrees of freedom is appropriate.

Using **standard error** estimate $\widehat{s.e.} = \frac{s}{\sqrt{n}} = \frac{0.082 \text{ ml/g/min}}{\sqrt{6}} = 0.03348$ ml/g/min yields

$$p\text{-value} = 2 P(\bar{X} \geq 0.767) = 2 P\left(T_5 \geq \frac{0.767 - 0.5}{0.03348}\right) = 2 P(T_5 \geq 7.976) = 2 (.00025) = .0005$$

This is strongly significant at any reasonable level α . According to the scale, the voxel should be assigned the color **RED**.

STATBOT 301, MODEL T

SUBJECT: BASIC CALCULATION OF P-VALUES FOR T-TEST

CALCULATE... from H_0

Test Statistic

$$\text{"t-score"} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Remember that the T-table corresponds to the area to the right of a positive t-score.



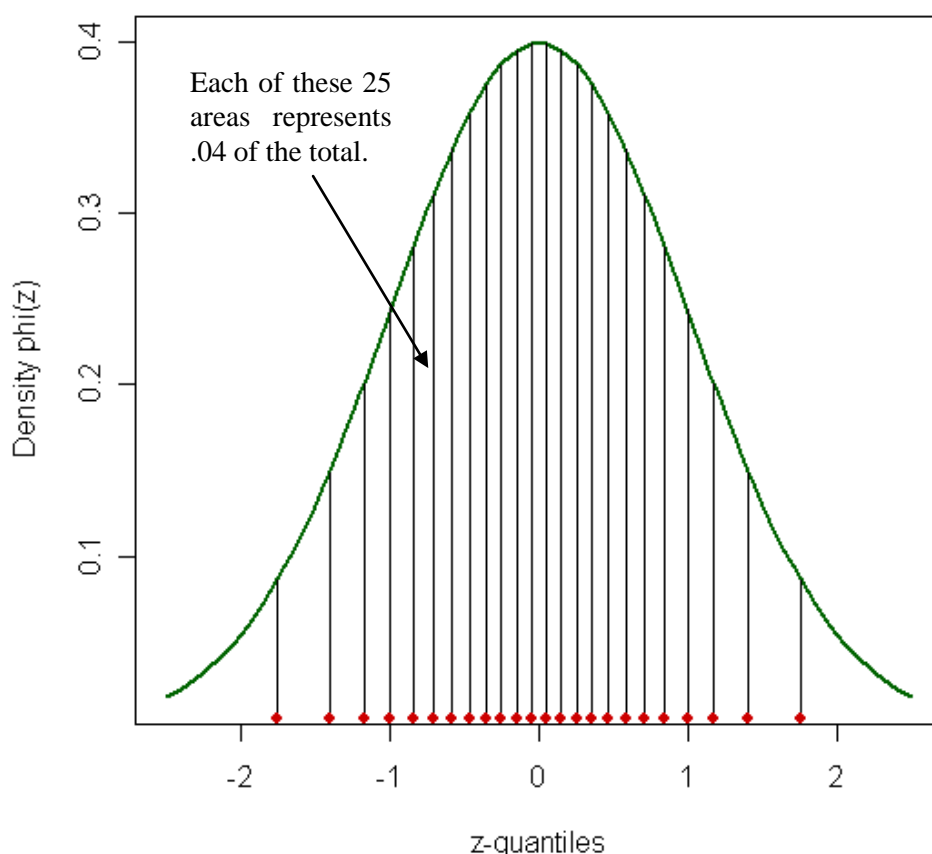
ALTERNATIVE HYPOTHESIS

	$H_A: \mu < \mu_0$	$H_A: \mu \neq \mu_0$	$H_A: \mu > \mu_0$
$t\text{-score}$ +	1 – table entry	2 × table entry	table entry
$t\text{-score}$ –	table entry for $ t\text{-score} $	2 × table entry for $ t\text{-score} $	1 – table entry for $ t\text{-score} $

Checks for normality ~ Is the ongoing assumption that the sample data come from a normally-distributed population reasonable?

- **Quantiles:** As we have already seen, $\approx 68\%$ within ± 1 s.d. of mean, $\approx 95\%$ within ± 2 s.d. of mean, $\approx 99.7\%$ within ± 3 s.d. of mean, etc. Other percentiles can also be checked informally, or more formally via...
- **Normal Scores Plot:** The graph of the quantiles of the n ordered (low-to-high) observations, versus the n known z -scores that divide the total area under $N(0, 1)$ equally (representing an ideal sample from the standard normal distribution), should resemble a straight line. Highly skewed data would generate a curved plot. Also known as a **probability plot** or **Q-Q plot** (for “Quantile-Quantile”), this is a popular method.

Example: Suppose $n = 24$ ages (years). Calculate the .04 quantiles of the sample, and plot them against the 24 *known* (i.e., “theoretical”) .04 quantiles of the standard normal distribution (below).



$\{-1.750, -1.405, -1.175, -0.994, -0.842, -0.706, -0.583, -0.468, -0.358, -0.253, -0.151, -0.050, +0.050, +0.151, +0.253, +0.358, +0.468, +0.583, +0.706, +0.842, +0.994, +1.175, +1.405, +1.750\}$

➤ Sample 1:

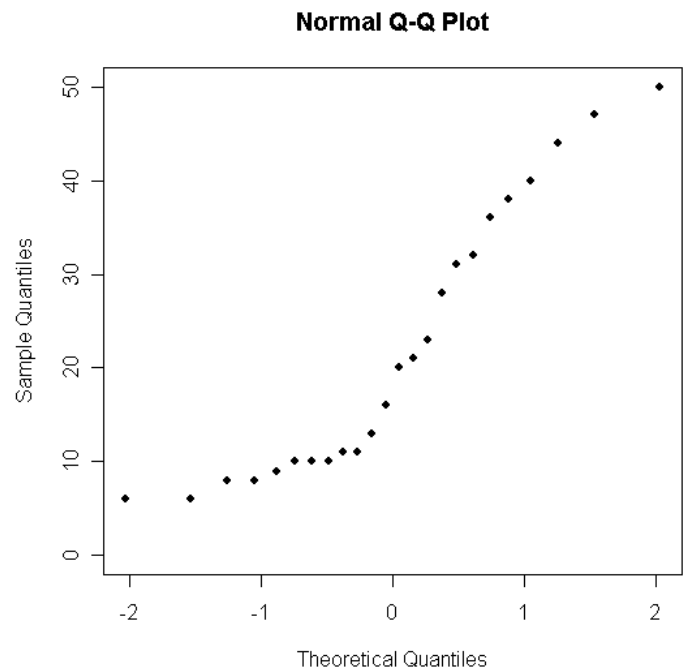
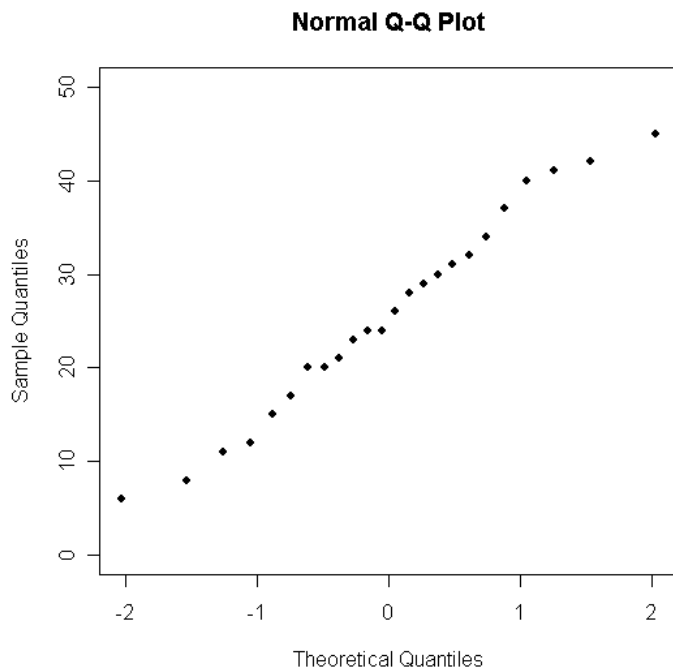
{6, 8, 11, 12, 15, 17, 20, 20, 21, 23, 24, 24, 26, 28, 29, 30, 31, 32, 34, 37, 40, 41, 42, 45}

The Q-Q plot of this sample (see first graph, below) reveals a more or less linear trend between the quantiles, which indicates that it is not unreasonable to assume that these data are derived from a population whose ages are indeed normally distributed.

➤ Sample 2:

{6, 6, 8, 8, 9, 10, 10, 10, 11, 11, 13, 16, 20, 21, 23, 28, 31, 32, 36, 38, 40, 44, 47, 50}

The Q-Q plot of this sample (see second graph, below) reveals an obvious deviation from normality. Moreover, the general “concave up” nonlinearity seems to suggest that the data are positively skewed (i.e., skewed to the right), and in fact, this is the case. Applying statistical tests that rely on the normality assumption to data sets that are not so distributed could very well yield erroneous results!



Formal tests for normality include:

- **Anderson-Darling**
- **Shapiro-Wilk**
- **Lilliefors** (a special case of **Kolmogorov-Smirnov**)

Remedies for non-normality ~ What can be done if the normality assumption is violated, or difficult to verify (as in a very small sample)?

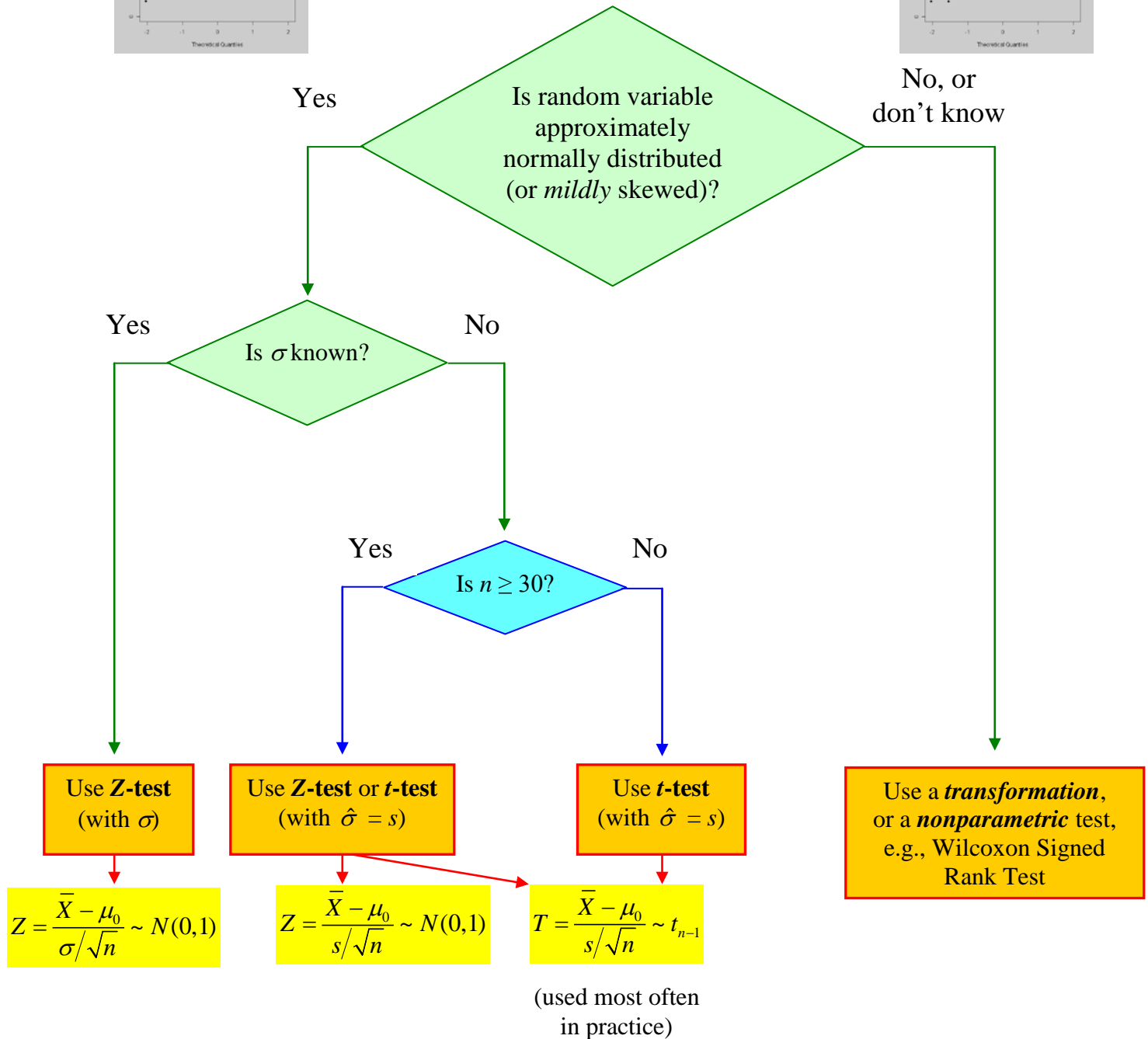
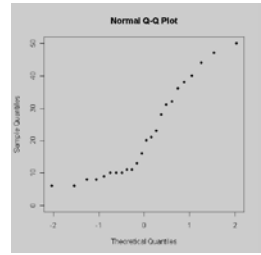
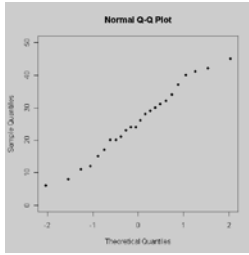
- **Transformations:** Functions such as $Y = \sqrt{X}$ or $Y = \ln(X)$, can transform a positively-skewed variable X into a normally distributed variable Y . (These functions “spread out” small values, and “squeeze together” large values. In the latter case, the original variable X is said to be *log-normal*.)

Exercise: Sketch separately the dotplot of X , and the dotplot of $Y = \ln(X)$ (to two decimal places), and compare.

X	$Y = \ln(X)$	Frequency
1		1
2		2
3		3
4		4
5		5
6		5
7		4
8		4
9		3
10		3
11		3
12		2
13		2
14		2
15		2
16		1
17		1
18		1
19		1
20		1

- **Nonparametric Tests:** Statistical tests (on the median, rather than the mean) that are free of any assumptions on the underlying distribution of the population random variable. Slightly less powerful than the corresponding parametric tests, tedious to carry out by hand, but their generality makes them very useful, especially for small samples where normality can be difficult to verify.

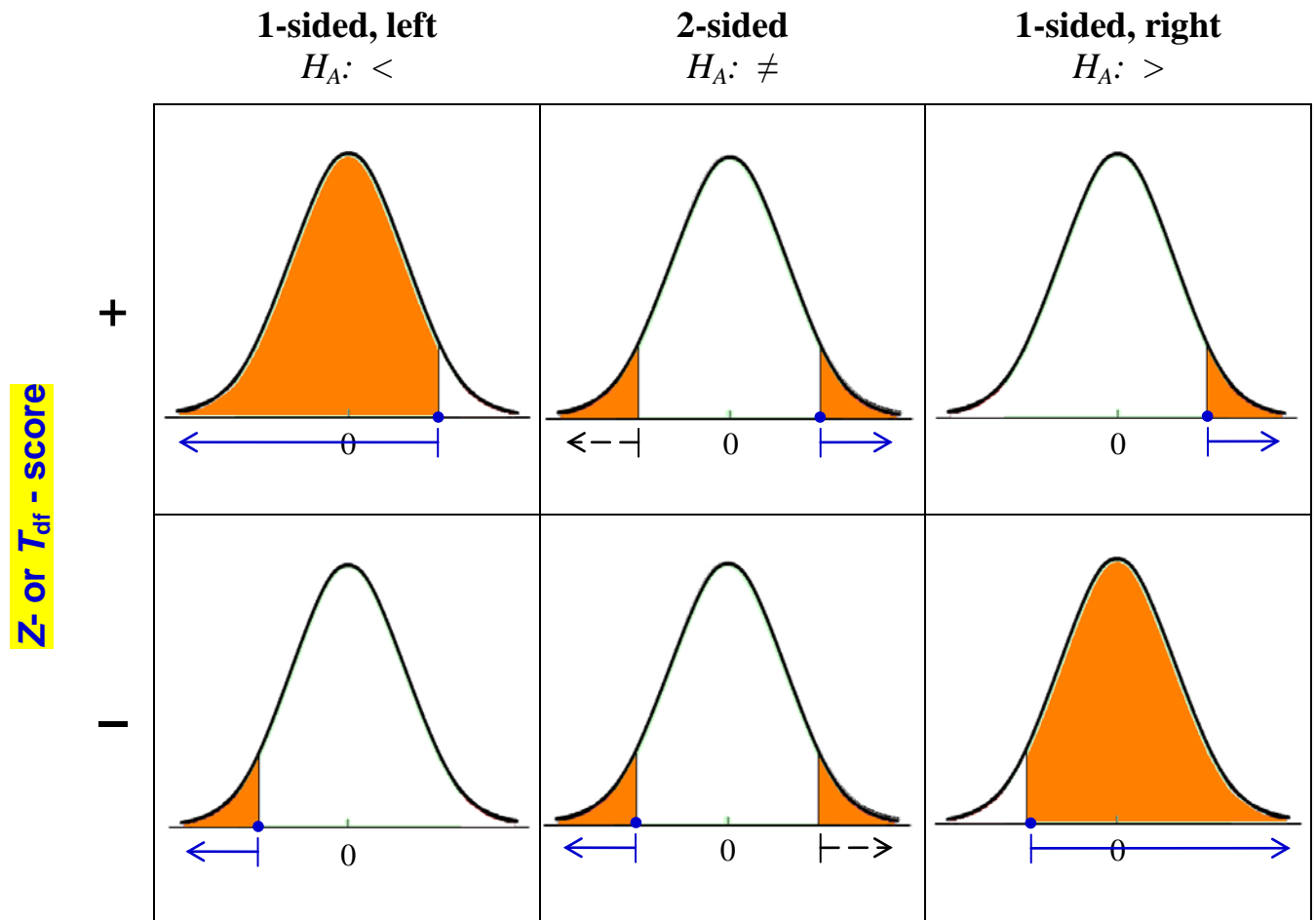
➤ **Sign Test** (crude), **Wilcoxon Signed Rank Test** (preferred)

GENERAL SUMMARY...**Step-by-Step Hypothesis Testing**
One Sample Mean $H_0: \mu \text{ vs. } \mu_0$ **CONTINUE...**

p-value: “How do I know in which direction to move, to find the **p-value?**”

See STATBOT, page 6.1-14 (Z) and page 6.1-24 (T), or...

Alternative Hypothesis

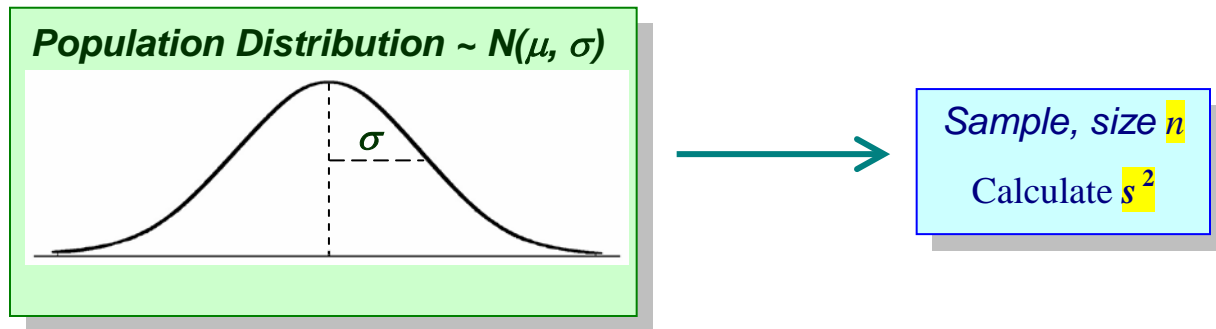


- The **p-value** of an experiment is the *probability* (hence always between 0 and 1) of obtaining a random sample with an outcome that is *as, or more*, extreme than the one actually obtained, *if* the null hypothesis is true.
- Starting from the value of the test statistic (i.e., z-score or t-score), the p-value is computed in the direction of the *alternative hypothesis* (either $<$, $>$, or both), which usually reflects the investigator's belief or suspicion, if any.
- If the p-value is “small,” then the sample data provides evidence that tends to *refute* the null hypothesis; in particular, if the p-value is *less* than the significance level α , then the null hypothesis can be *rejected*, and the result is **statistically significant** at that level. However, if the p-value is *greater* than α , then the null hypothesis is *retained*; the result is not statistically significant at that level. Furthermore, if the p-value is “large” (i.e., close to 1), then the sample data actually provides evidence that tends to *support* the null hypothesis.

§ 6.1.2 Variance

Given: Null Hypothesis $H_0: \sigma^2 = \sigma_0^2$ (constant value)

versus Alternative Hypothesis $H_A: \sigma^2 \neq \sigma_0^2$ Two-sided Alternative
Either $\sigma^2 < \sigma_0^2$ or $\sigma^2 > \sigma_0^2$

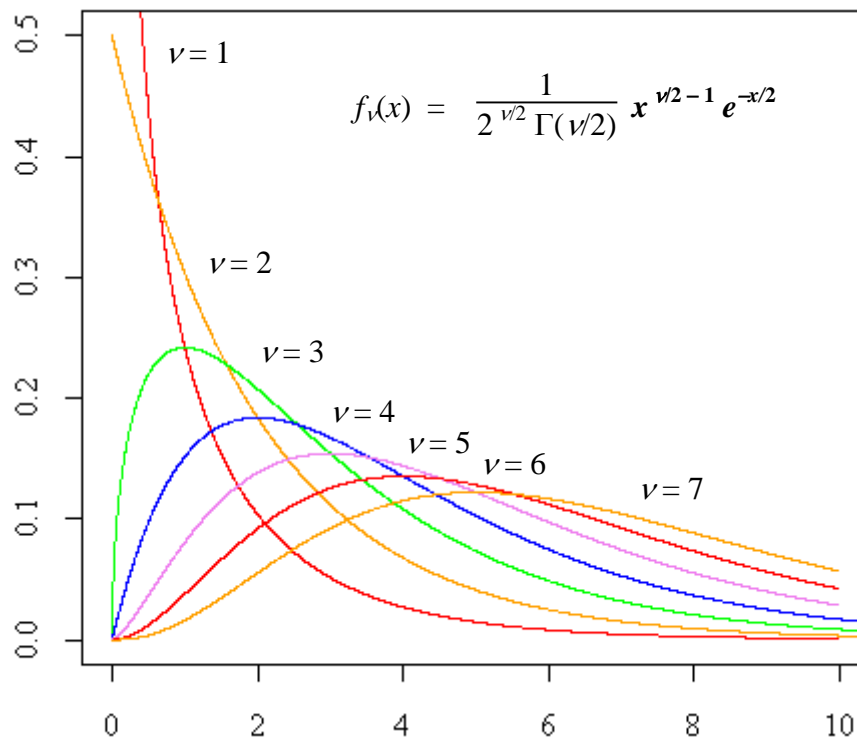


Test statistic:

$$X^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Sampling Distribution of X^2 :

Chi-Squared Distribution, with $\nu = n - 1$ degrees of freedom $df = 1, 2, 3, \dots$



Note that the chi-squared distribution is not symmetric, but *skewed to the right*. We will not pursue the details for finding an acceptance region and confidence intervals for σ^2 here. But this distribution will appear again, in the context of hypothesis testing for equal proportions.

§ 6.1.3 Proportion

POPULATION

Binary random variable

$$Y = \begin{cases} 1, & \text{Success with probability } \pi \\ 0, & \text{Failure with probability } 1 - \pi \end{cases}$$

↓

Experiment: n independent trials

↓

SAMPLE

Random Variable: $X = \# \text{ Successes} \sim \text{Bin}(n, \pi)$

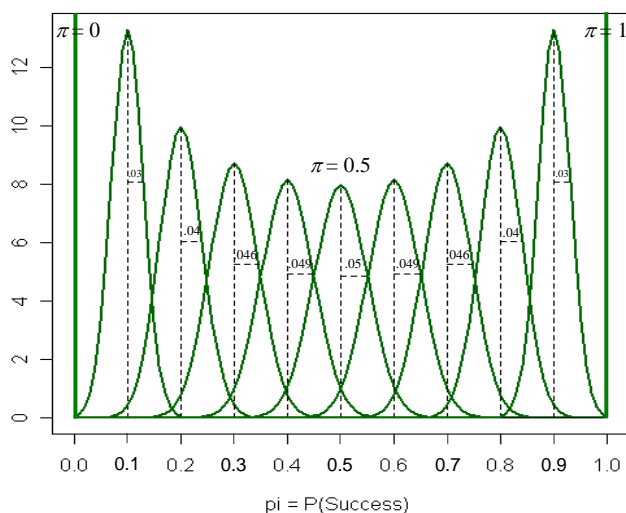
Recall: Assuming $n \geq 30$, $n\pi \geq 15$, and $n(1 - \pi) \geq 15$,

$$X \sim N(n\pi, \sqrt{n\pi(1 - \pi)}), \text{ approximately. (see §4.2)}$$

Therefore, dividing by n ...

$$\hat{\pi} = \frac{X}{n} \sim N\left(\pi, \underbrace{\sqrt{\frac{\pi(1 - \pi)}{n}}}_{\text{standard error s.e.}}\right), \text{ approximately.}$$

Problem! The expression for the standard error involves the very parameter π upon which we are performing statistical inference. (This did not happen with inference on the mean μ , where the standard error is $\text{s.e.} = \sigma / \sqrt{n}$, which does not depend on μ .)



← Illustration of the bell curves $N\left(\pi, \sqrt{\frac{\pi(1 - \pi)}{n}}\right)$

for $n = 100$, as proportion π ranges from 0 to 1. Note how, rather than being fixed at a constant value, the “spread” s.e. is smallest when π is close to 0 or 1 (i.e., when success in the population is either very rare or very common), and is maximum when $\pi = 0.5$ (i.e., when both success and failure are equally likely).

Also see Problem 4.4/10. This property of nonconstant variance has further implications; see “Logistic Regression” in section 7.3.

Example: Refer back to the coin toss example of section 1.1, where a random sample of $n = 100$ independent trials is performed in order to acquire information about the probability $P(\text{Heads}) = \pi$. Suppose that $X = 64$ Heads are obtained. Then the sample-based **point estimate** of π is calculated as $\hat{\pi} = X / n = 64/100 = 0.64$. To improve this to an **interval estimate**, we can compute the...

$(1 - \alpha) \times 100\%$ Confidence Interval for π

$$\left(\hat{\pi} - z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}, \hat{\pi} + z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \right)$$

95% Confidence Interval for π

$$95\% \text{ limits} = 0.64 \pm z_{0.025} \sqrt{\frac{(0.64)(0.36)}{100}} = 0.64 \pm 1.96 (.048) \quad \leftarrow \widehat{\text{s.e.}} = .048$$

$\therefore 95\% \text{ CI} = (0.546, 0.734)$ contains the true value of π , with 95% confidence.

Is the coin fair at the $\alpha = .05$ level?

Null Hypothesis $H_0: \pi = 0.5$

vs. **Alternative Hypothesis** $H_A: \pi \neq 0.5$

\neq

As the 95% CI does not contain the null-value $\pi = 0.5$, H_0 can be rejected at the $\alpha = .05$ level, i.e., the coin is not fair.

$(1 - \alpha) \times 100\%$ Acceptance Region for $H_0: \pi = \pi_0$

$$\left(\pi_0 - z_{\alpha/2} \sqrt{\frac{\pi_0(1 - \pi_0)}{n}}, \pi_0 + z_{\alpha/2} \sqrt{\frac{\pi_0(1 - \pi_0)}{n}} \right)$$

95% Acceptance Region for $H_0: \pi = 0.50$


$$95\% \text{ limits} = 0.50 \pm z_{0.025} \sqrt{\frac{(0.50)(0.50)}{100}} = 0.50 \pm 1.96 (.050) \quad \leftarrow \text{s.e.}_0 = .050$$

$\therefore 95\% \text{ AR} = (0.402, 0.598)$

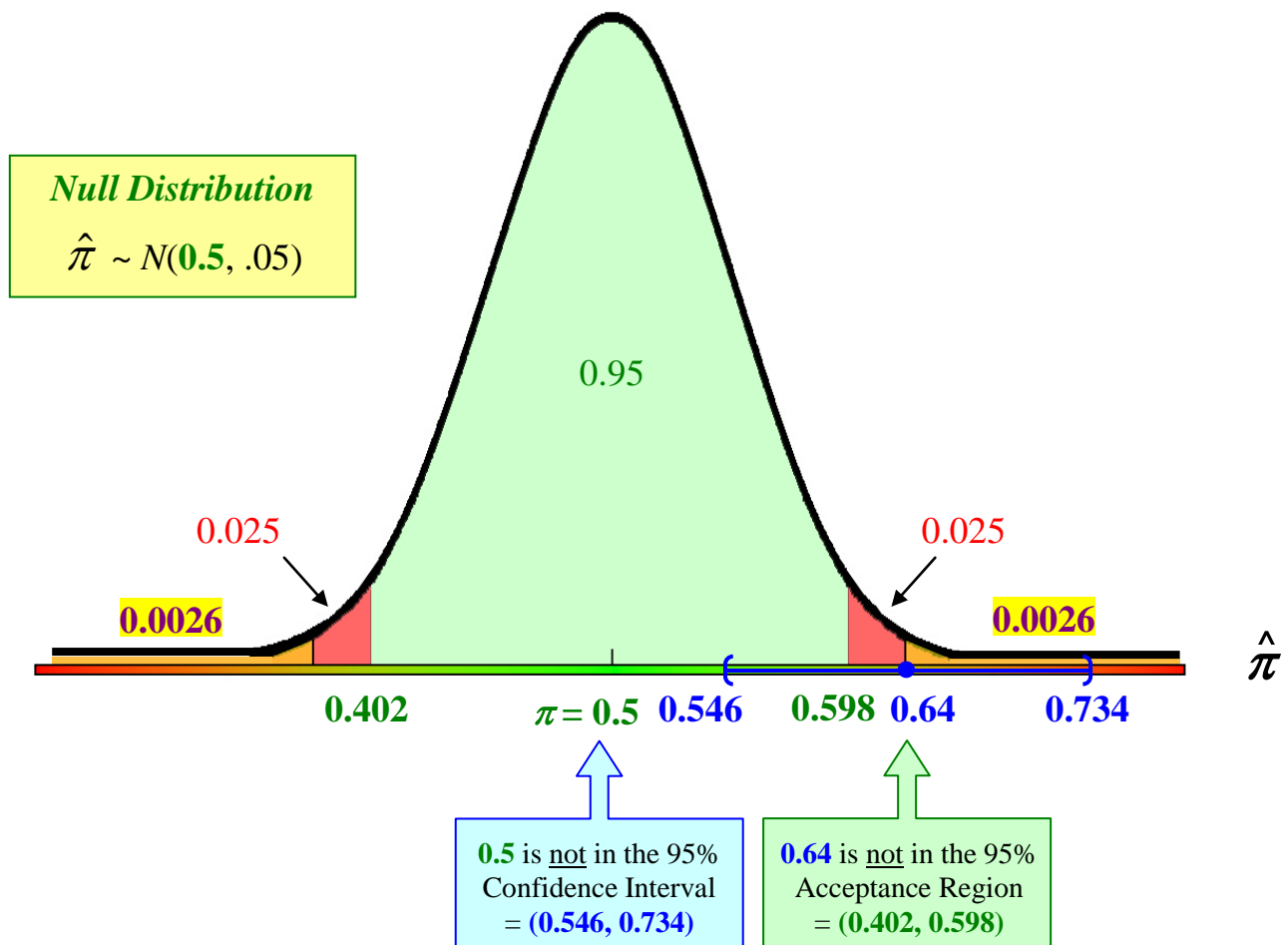
As the 95% AR does not contain the sample proportion $\hat{\pi} = 0.64$, H_0 can be rejected at the $\alpha = .05$ level, i.e., the coin is not fair.

Test Statistic

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}} \sim N(0, 1)$$

 **p-value** = $2 P(\hat{\pi} \geq 0.64) = 2 P\left(Z \geq \frac{0.64 - 0.50}{.050}\right) = 2 P(Z \geq 2.8) = 2(.0026) = .0052$

As $p \ll \alpha = .05$, H_0 can be strongly rejected at this level, i.e., the coin is not fair.



Comments:

- A **continuity correction** factor of $\pm \frac{0.5}{n}$ may be added to the numerator of the Z test statistic above, in accordance with the “normal approximation to the binomial distribution” – see 4.2 of these Lecture Notes. (The “ n ” in the denominator is there because we are here dealing with *proportion* of success $\hat{\pi} = X / n$, rather than just *number* of successes X .)
- Power and sample size calculations are similar to those of inference for the mean, and will not be pursued here.

See [Appendix > Statistical Inference > General Parameters and **IMPORTANT FORMULA TABLES**](#).
and [Appendix > Statistical Inference > Means and Proportions, One and Two Samples](#).