Prelim 2 for BTRY6010/ILRST6100

November 1, 2018: 7:30pm-9:30pm

Name:_____

NetID:_____

Lab: (circle one)

Lab 402: Tues 1:25PM - 2:40PM

Lab 403: Tues 2:55PM - 4:10PM

Lab 404: Wed 2:55PM - 4:10PM

Lab 405: Tues 7:30PM - 8:45PM

Score: _____ / 65

Instructions

1. Please **do not turn to next page** until instructed to do so.
2. You have 120 minutes to complete this exam.
3. The last two pages of this exam have some useful formulas.
4. No textbook, calculators, phone, computer, notes, etc. allowed (please keep your phones off or do not bring them to the exam).
5. Please answer questions in the spaces provided.
6. When asked to calculate a number, it is sufficient to write out the full expression in numbers or code without actually calculating the value. E.g., $\frac{1+3\times\frac{4}{7}}{3+0.7}$ or `1 - pnorm(3)^2` are valid answers.
7. Please read the following statement and sign before beginning the exam.

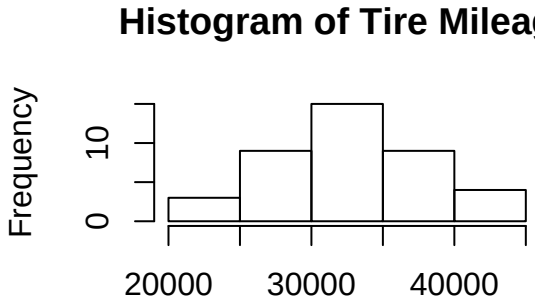
Academic Integrity

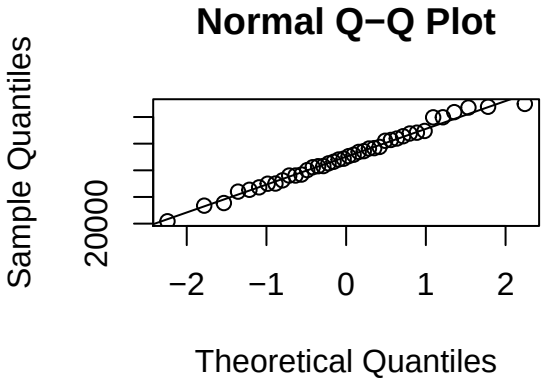
I, _____, certify that this work is entirely my own. I will not look at any of my peers' answers or communicate in any way with my peers. I will not use any resource other than a pen/pencil. I will behave honorably in all ways and in accordance with Cornell's Code of Academic Integrity.

Signature: _____ **Date:**_____

For the following question an answer without justification will not receive full credit.

1. A factory makes 2,000 tires a day, and mileage of tires (denoted by a random variable X) are independent of each other. One day, a simple random sample of 40 tires was drawn and their mileages were tested. It was found that the sample standard deviation of their mileage was 5347.58 miles. Figures and summary statistics of the mileages of the 40 sample tires are shown below.





minimum	1st quartile	median	mean	3rd quartile	maximum
20444.65	29020.58	32420.57	32590.95	36069.02	42454.10

We are interested in the mean mileage of tires made on that day in the factory.

(a) [1 point] Define the parameter of interest (μ) in the context of this problem.

ANSWER: average mileage (or population mean mileage) of all 2000 tires made on that day in the factory.

(b) [3 points] Give a point estimator of μ and specify its distribution. What is the realized value of your point estimator in this sample? [Hint: point estimator is a random variable, its realization is a numeric value).

ANSWER

- Point estimator of μ and its distribution:

$$\bar{x}_{40} = \frac{1}{40}\sum_{i=1}^{40} x_i \sim N\left(\mu, \frac{\sigma}{\sqrt{40}}\right)$$

where σ is the standard deviation of the mileage of all 2000 tires made on that day

OR

$$\bar{x}_{40} = \frac{1}{40}\sum_{i=1}^{40} x_i \sim N\left(\mu, \frac{5347.58}{\sqrt{40}}\right) \quad \text{approximately}$$

Note: μ is unknown but fixed; you cannot substitute sample mean $\bar{x}_{40}$ for population mean μ in the distribution.

(c) [3 points] Clearly state and/or check any assumption you made to arrive at the conclusion in (b).

ANSWER

- **Independence**: The mileage of tires made on that day are independent of each other.

  AND either one of the following (If you use t-distribution for confidence interval and/or hypothesis testing, then you must include **Normality** condition. Otherwise, you can include either **Normality** or **Sample size** condition.)

- **Normality**: From the histogram and the Q-Q plot, we see that the mileage of tires made on that day is approximately normal.

  OR

- **Sample size**: The sample size $n = 40 > 30$; the sample size is sufficiently large for Central Limit Theorem to apply.

(d) [4 points] Based on this sample of 40 tires, construct a 95% confidence interval for μ, and interpret it in the context of the problem.

ANSWER (using normal quantile):

- $1 - \alpha = 95\%$ so that $\alpha = 0.05$; $z_{\alpha/2} =$ `-qnorm(0.05/2)` $= 1.96$; $SE(\bar{X}) = s_{40}/\sqrt{40} = 5347.58/\sqrt{40}$

$$\begin{aligned}&(\bar{x}_{40} - z_{\alpha/2}SE(\bar{X}_n), \bar{x}_{40} + z_{\alpha/2}SE(\bar{X}_n))\\ =&(32590.95 - z_{0.05/2} \times 5347.58/\sqrt{40}, 32590.95 + z_{0.05/2} \times 5347.58/\sqrt{40})\\ =&(30933.72, 34248.18) \quad \text{(You don't need to compute this)}\end{aligned}$$

- We are 95% confident that the average mileage of all 2000 tires made on that day in the factory is between $32590.95 - z_{0.05/2} \times 5347.58/\sqrt{40}$ and $32590.95 + z_{0.05/2} \times 5347.58/\sqrt{40}$.

ANSWER (using t quantile)

- $1 - \alpha = 95\%$ so that $\alpha = 0.05$; $t_{\alpha/2,n-1} =$ `-qt(0.05/2, df = 40-1)` $= 2.02$; $SE(\bar{X}) = s_{40}/\sqrt{40} = 5347.58/\sqrt{40}$

$$\begin{aligned}&(\bar{x}_{40} - z_{\alpha/2}SE(\bar{X}_n), \bar{x}_{40} + z_{\alpha/2}SE(\bar{X}_n))\\ =&(32590.95 - t_{\alpha/2,n-1} \times 5347.58/\sqrt{40}, 32590.95 + t_{\alpha/2,n-1} \times 5347.58/\sqrt{40})\end{aligned}$$

- We are 95% confident that the average mileage of all 2000 tires made on that day in the factory is between $32590.95 - t_{\alpha/2,n-1} \times 5347.58/\sqrt{40}$ and $32590.95 + t_{\alpha/2,n-1} \times 5347.58/\sqrt{40}$.

Now, suppose we want to test whether or not the mean mileage of tires made on that day is $30,000$ miles, with a significance level of 5%.

(e) [3 points] What are the null and alternative hypotheses? Interpret the significance level $\alpha = 5\%$ in the context of the problem.

ANSWER

- Null hypothesis

$$H_0 : \mu = 30,000$$

- Alternative hypothesis

$$H_A : \mu \neq 30,000$$

- When the average mileage of the 2,000 tires made on that day is indeed 30,000 miles (i.e., given H_0 is true), there is a probability of 5% for the test that you pick to reject the null hypothesis and claim that the average mileage is NOT 30,000 miles (i.e., the probability of incorrectly rejecting H_0).

(f) [3 points] What is the test statistic (expressed as a random variable) and what is its distribution under the null hypothesis?

ANSWER

Three versions of test statistics

- $T = \frac{\bar{x}_{40} - 30000}{5347.58/\sqrt{40}} \sim t_{39}$ under H_0
- $Z = \frac{\bar{x}_{40} - 30000}{5347.58/\sqrt{40}} \sim N(0,1)$ under H_0
- $\bar{x}_{40} \sim N(30000, \frac{5347.58}{\sqrt{40}})$ under H_0

Note 1: you cannot have the unknown σ inside any part of your test statistic and its distribution.

Note 2: a realized value of test statistic does not follow a distribution.

(g) [2 points] Calculate the realized value of the test statistic in this example.

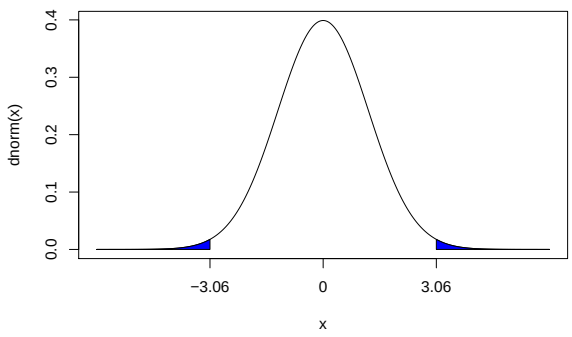
ANSWER

- $T = \frac{\bar{x}_{40} - 30000}{5347.58/\sqrt{40}} \sim t_{39}$

(h) [3 points] What is the p-value of this test? (You should draw a picture as you answer this, clearly mark the area used to calculate p-value, and provide an R command you would use to calculate this.)

ANSWER

p-value for the test is `2(1-pnorm(3.0643))` = 0.0021818.



Note: can also be computed using t-distribution or $N(30000, \frac{5347.58}{\sqrt{40}})$ if you use $T = \frac{\bar{x}_{40} - 30000}{5347.58/\sqrt{40}}$ or $\bar{x}_{40}$, respectively, as your test statistic in question (f).

(i) [2 points] Clearly write down the assumptions you would check to see if the distribution of test statistic you stated before is acceptable.

ANSWER

- **Independence**: The mileage of tires made on that day are independent of each other.

  AND either one of the following (If you use t-distribution for confidence interval and/or hypothesis testing, then you must include **Normality** condition. Otherwise, you can include either **Normality** or **Sample size** condition.)

- **Normality**: From the histogram and the Q-Q plot, we see that the mileage of tires made on that day is approximately normal.

  OR

- **Sample size**: The sample size $n = 40 > 30$; the sample size is sufficiently large for Central Limit Theorem to apply.

These assumptions are essentially the same with those in (b).

(j) [2 points] Let us assume that the p-value of this test is smaller than 0.03. Based on this, is there sufficient evidence in the data to say that the mean mileage of tires made on that day is different from $30,000$ miles?

ANSWER

The p-value $< 0.03 < 0.05$, the significance level. Since the p-value for this sample is smaller than the significance level of this test, we reject the null hypothesis ($\mu = 30000$) in favour of the alternative ($\mu \neq 30000$). Thus it is justified to say that the mean life of the tires that were made on that day is different from 30,000 miles.

(k) [2 points] Comment on whether we could have known the decision of this hypothesis test (reject or fail to reject null) based on the 95% confidence interval constructed before.

ANSWER Yes, we could have because the confidence interval didn not include the null hypothesis of 30,000 miles coming from the tires.

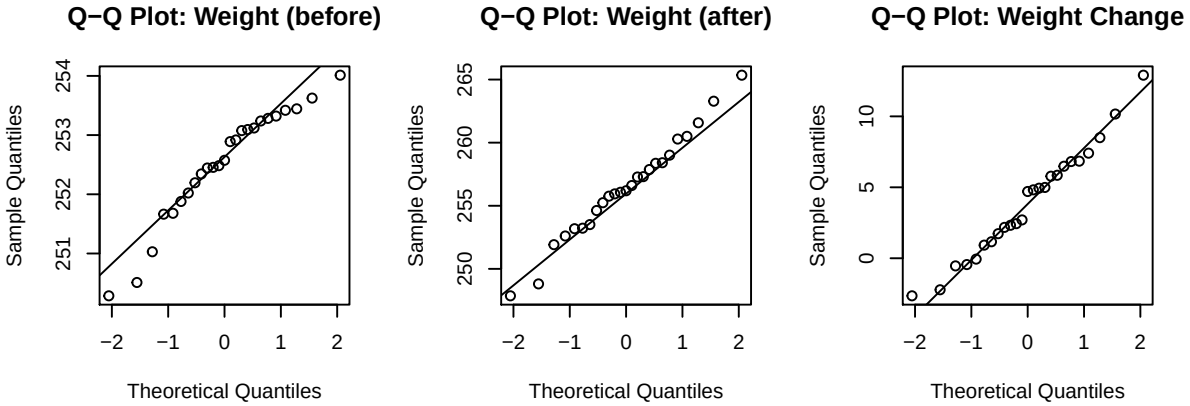
(Also acceptable): if the confidence interval contains 30,000, we could have known that the decision of this hypothesis will be 'fail to reject H_0', and vice versa.

(l) [2 points] Based on this p-value and the summary statistics, is it possible to know the decision of the hypothesis test $H_0 : \mu = 30,000$ vs. $H_A : \mu < 30,000$ at 5% level of significance?

ANSWER The summary statistic table shows that sample mean is $> 30,000$. So the p-value of this test will be larger than 0.5, so we will fail to reject H_0.

For the following question an answer without justification will not receive full credit.

2. We administered a drug believed to affect weight to $n = 25$ rats. The rats were weighed immediately before administration of the drug, and after one month, the rats were weighed again. Before, the mean rat weight was 252.5 ounces. One month later, the mean rat weight was 256.5 ounces. The sample standard deviation of the weight change was 5 ounces. The following quantile-quantile plots suggest approximate normality of rat weights after 1 month and their weight changes.



We are interested in seeing if, on average, there was an *increase* in rat weight one month after drug administration.

(a) [1 point] Define the parameter of interest (μ_d) in the context of this problem.

μ_d is the population mean difference in the weight of the rats before and one month after the drug was administered (after - before).

(b) [4 points] Specify the distribution of your sample statistic, and calculate its realized value in this data. Clearly state and/or check any assumption you made to arrive at this conclusion.

With population mean μ_d and population standard deviation σ_d, the sample statistic $\overline{X}_d \sim N(\mu_d, \frac{\sigma_d}{\sqrt{25}})$. We are assuming that the changes in rat weight are normally distributed (which looks reasonable based on the Q-Q Plot for weight change), and independent. The realization of the sample statistic is $\overline{x}_d = 256.2 - 252.2 = 4$

(c) [4 points] Construct a 90% confidence interval for $\hat{\mu}_d$, and interpret it in the context of the problem.

Since the samples size is small and the population standard deviation is unknown, we will use a t quantile in calculating the confidence interval. We have $\alpha = 0.1$ and 24 degrees of freedom, so we need $\overline{x}_d \pm t_{24,0.05}\frac{s_d}{\sqrt{25}}$. Plugging in, that is $4 \pm t_{24,0.05}\frac{5}{\sqrt{25}}$, which can be simplified to $4 \pm t_{24,0.05}$. We can calculate $t_{24,0.05}$ in R with the command `-qt(0.05,24)`.

The interpretation is that we are 90% confident that the population mean difference in rat weights is in the between $4 - t_{24,0.05}$ and $4 + t_{24,0.05}$.

(d) [2 points] Looking at your confidence interval, your colleague had the impression that there is a 90% probability that the true increase in average weight is somewhere in your confidence interval. Please provide him with an explanation (in words) of what "90% confidence level" means in the context of your problem.

If we were to repeat the experiment and calculate a 90% confidence interval each time, then on average 90% of those intervals will contain the population mean difference in rat weight.

(e) [2 points] Comment on how you would expect the length of the 90% confidence interval to change if we had only $n = 4$ rats in our sample. Assume the sample standard deviation is still 5 ounces.

The length of the confidence interval is determined by $t_{n-a,\alpha/2}\frac{s_d}{\sqrt{n}}$.

If $n = 4$, $\frac{s_d}{\sqrt{n}}$would increase from $\frac{5}{\sqrt{25}}$ to $\frac{5}{\sqrt{4}}$

and with fewer degrees of freedom, $t_{3,\alpha/2} > t_{24,\alpha/2}$as well. so the length of the CI will increase.

(f) [1 point] Specify all the Q-Q plots that you used to justify normality assumption (if any) for your answer in (b).

Only plot 3 is needed since we are dealing with the weight change.

Now, suppose that we want to test if there was an *increase* in the weight after drug administration, with a significance level of 0.5%.

(g) [2 points] What are the null and alternative hypotheses?

$$H_0 : \mu_d = 0$$
$$H_A : \mu_d > 0$$

(h) [2 points] What is the test statistic (expressed as a random variable) and what is its distribution under the null hypothesis?

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} \sim t_{24}$$

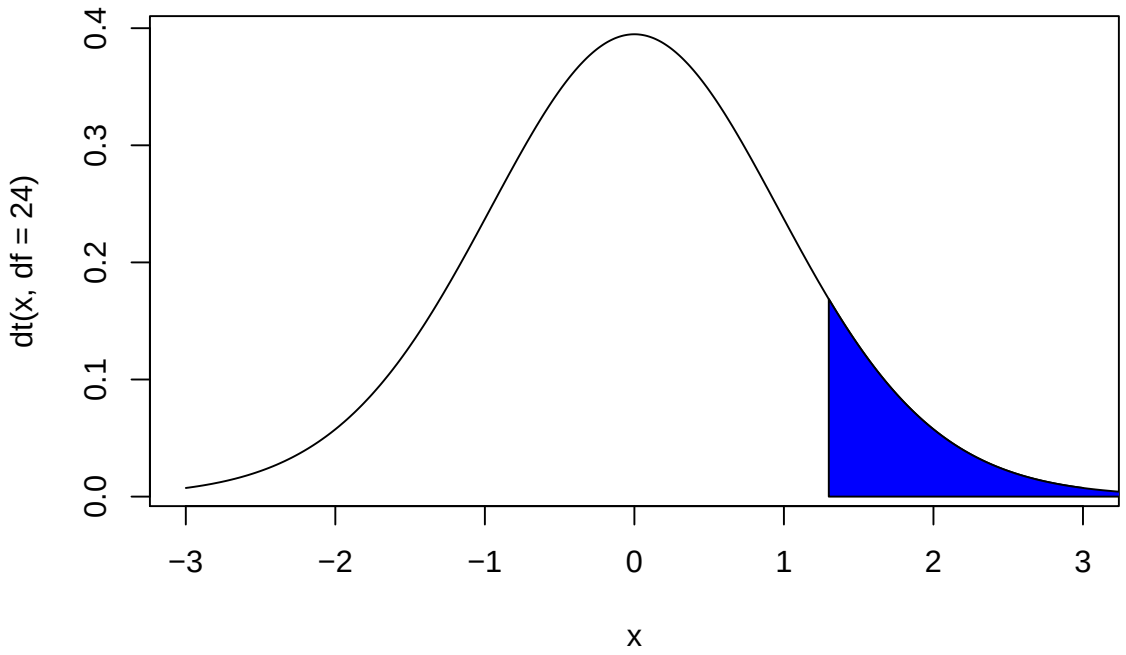
(i) [1 point] Calculate the realized value of the test statistic in this example.

$$t = \frac{256 - 252 - 0}{5/\sqrt{25}} = 4$$

(j) [2 points] What is the p-value of this test? (You should draw a picture as you answer this, clearly mark the area used to calculate p-value, and provide an R command you would use to calculate this.)

$$1 - pt(tstat, df = 24)$$

```
z = 1.3
x = seq(-3, 3, length.out = 1000)
{plot(x, dt(x,df=24), type = 'l')
y = seq(z, 4, length.out = 1000)
polygon(c(z,y,4), c(0,dt(y,df=24),0), col = "blue") }
```



(k) [2 points] Assume the p-value is less than 0.01. Based on this, report the conclusion of your test in words.

Since it is not known if the p-value is less than 0.05%, we do not have enough information to conclude at 0.05% level of significance whether the population mean of rat weight gain is not zero.

Also acceptable due to clarity: any correct interpretation for a specific α.

(l) [2 points] Comment on what information in your data (other than the p-value) can help a scientist decide if the results are *practically* significant.

The effect size should be considered. In this case, the sample mean of weight gain in rats could provide information on whether this change in weight is substantial enough to conclude that the drug is effective.

(m) [2 points] Explain in words what "power of a test" means in the context of this problem. What can be done to increase power of this test when significance level is set at 1%?

$$\text{definition of power} : P(\text{reject } H_0 | \mu_d > 0), \text{probability of missing a real change in weight}$$

Increase n the number of samples in trial.

(n) [2 points] Suppose we would like to test if, on average, there was a *decrease* in rat weight between drug administration and one month later. A new set of data is collected, where the sample means of weights before and after drug administration are 252 and 255 ounces, respectively. What can be said about the p-value of this test?

This p-value will be larger (greater than 0.5) because the test statistic is now positive($\frac{255-252}{5/\sqrt{25}} = 3$)

and the alternative is $H_A : \mu_d < 0$.

3. [6 points] The following code attempts to generate a sampling distribution of $\bar{X}_4 = (X_1+X_2+X_3+X_4)/4$, where X_1, X_2, X_3, X_4 are independent normal random variables with mean 50 and *variance* 4. Then it attemps to overlay the true density function of $\bar{X}_4$ on top of the probability histogram obtained from the simulated data.

```
#set n
n = 4


#initialize realizations
num.simulations = 10000


xbar.realizations = rep(NA, num.simulations)



#simulate realizations
for (i in 1:num.simulations) {


  xbar.realizations[i] = mean(4*rnorm(n)+50) # <---- incorrect
  xbar.realizations[i] = mean(2*rnorm(n)+50) # <---- correct


}


#plot histogram
hist(xbar.realizations, # <---- incorrect
hist(xbar.realizations, freq = FALSE  # <--- correct
main = expression(paste('Histogram of 10,000 Draws of ',bar(X)[n])))



#overlay true density
xvalues = seq(47, 53, 0.01)


yvalues = dnorm(xvalues,50, sqrt(2/n)) # <---- incorrect
yvalues = dnorm(xvalues,50, sqrt(4/n)) # <---- correct


lines(xvalues,yvalues)
```

Now this code has some error/omission in it. Your job is to fix that. At the end of your corrections we should have a code chunk such that running it in `R` would give us the correct graphs.

If you think a line of code is correct then leave it as is. If you think a line of code has error(s) or omission(s) in it, then scratch it out and write the correct code in the space provided directly below that line.

Formula Sheet

The law of total probability:

$$P(B) = P(A)P(B|A) + P(A^C)P(B|A^C)$$

Bayes' theorem:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(A^C)P(B|A^C)}.$$

Binomial distribution: $X \sim \text{Binomial}(n, p)$

- probability mass function (pmf):

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n$$

- expected value $E(X) = np$
- variance $\text{Var}(X) = np(1-p)$

Poisson distribution: $X \sim \text{Poisson}(\lambda)$

- probability mass function (pmf):

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \ldots$$

- expected value $E(X) = \lambda$
- variance $\text{Var}(X) = \lambda$

Uniform distribution: $X \sim Unif(a, b)$

- probability density function (pdf):

$$f(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$

- expected value $E(X) = (a+b)/2$
- variance $\text{Var}(X) = (b-a)^2/12$

Normal distribution: $X \sim \text{Normal}(\mu, \sigma)$

- probability density function (pdf):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty$$

- expected value $E(X) = \mu$
- variance $\text{Var}(X) = \sigma^2$
- If $X \sim N(\mu, \sigma)$, then $Z = \frac{X-\mu}{\sigma}$ has a $N(0, 1)$ distribution.

Scenario	**One population Mean**	**One population proportion**	**Paired Mean**
Parameter	μ	p	μ_D
Statistic / Point Estimate	$\bar{x}$	$\hat{p} = \frac{X}{n}$	$\bar{x}_D$
Standard Error (of point estimate)	$SE(\bar{x}) = \sigma/\sqrt{n}$ (σ known) $SE(\bar{x}) = S/\sqrt{n}$ (σ unknown)	$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE(\bar{x}_D) = \sigma_D/\sqrt{n}$ (σ_D known) $SE(\bar{x}_D) = S_D/\sqrt{n}$ (σ unknown)
Confidence Interval	$(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ (σ known) $(\bar{x} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}})$ (σ unknown)	$\hat{p} \pm z_{\alpha/2}SE(\hat{p})$	$(\bar{x}_D \pm z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}})$ (σ_D known) $(\bar{x}_D \pm t_{n-1,\alpha/2}\frac{S_D}{\sqrt{n}})$ (σ_D unknown)
Hypothesis Testing	$Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ (σ known) $T = \frac{\bar{x}-\mu_0}{S/\sqrt{n}}$ (σ unknown)	$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$Z = \frac{\bar{x}_D-0}{\sigma_D/\sqrt{n}}$ (σ_D known) $T = \frac{\bar{x}_D-0}{S_D/\sqrt{n}}$ (σ_D unknown)

- **Sampling Distribution of** $\bar{X}$: If X has expectation μ and standard deviation σ. Then, $E(\bar{X}) = \mu$, $SD(\bar{X}) = \sigma/\sqrt{n}$.
  - If $X \sim N(\mu, \sigma)$, then $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$
  - For large sample ($n \geq 30$), under suitable assumptions on X, $\bar{X}$ is *approximately* $N(\mu, \sigma/\sqrt{n})$
- **Sampling Distribution of** $\hat{p}$: $\hat{p} = X/n$, $E(\hat{p}) = p$, $SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.
  - For large sample ($np \geq 10, \quad n(1-p) \geq 10$), $\hat{p}$ is *approximately* $N(p, \sqrt{\frac{p(1-p)}{n}})$.
- t**-distribution**: If $X \sim N(\mu, \sigma)$, then $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a t-distribution with degree of freedom $n-1$.
  - With larger n, t-distribution with n degrees of freedom becomes more similar to $N(0,1)$.