

Thought: *Never do card tricks for the people you play poker with.*

Thought: *To succeed in politics, it is often necessary to rise above your principles.*

Monday 11/4/02 : OPTIONAL REVIEW, ask questions about homework, course material, sample exam, etc. ***Suggestion - print out Sample Exam 2 and bring it to class with you.***

Help for Exam 2

- Monday, periods 3–8, FLO 104
- Monday, 6:15 pm – 8:10 pm, McCC 100

Tuesday 11/5/02 : EXAM 2—during your regularly scheduled

DISCUSSION SECTION

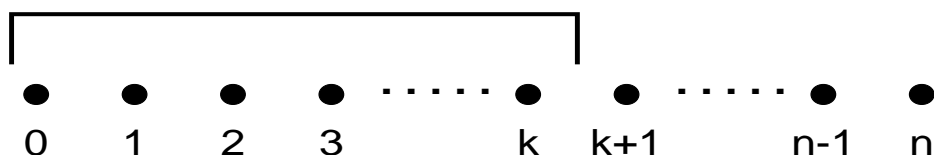
Wednesday 11/6/02 : Minitab, Computer Demo,
P. 288 – 393 (Sec. 7.2), P. 341 – 345 (Sec. 8.4)

Chapter 4 : Discrete Random Variables

- Can COUNT the number of distinct values of the variable.
- The Binomial Probability Distribution
 - Some discrete random variables are binomial – NOT ALL
 - Binomial Experiment p. 179 – 5 criteria
 - If $n =$ number of trials, $P(S) = p$: (p. 183)

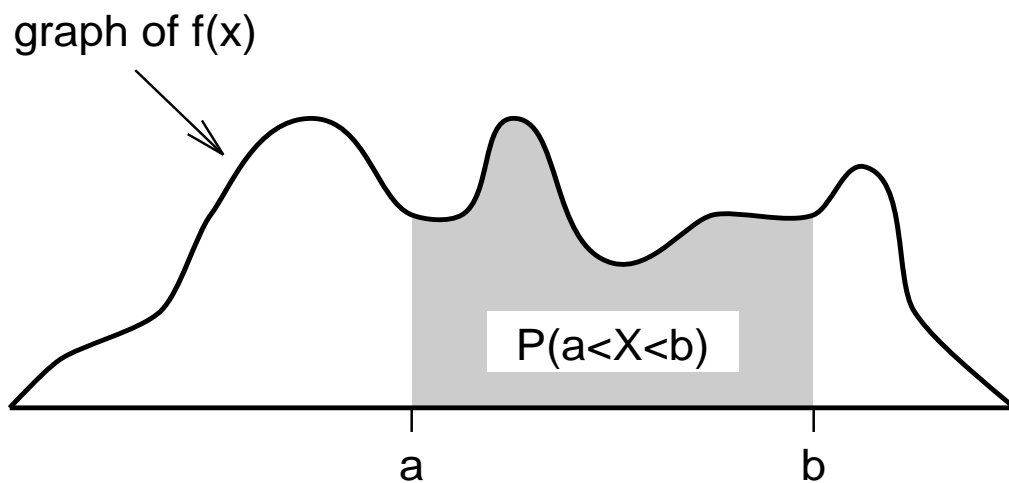
$$p(x) = \binom{n}{x} p^x q^{n-x} \text{ for } x = 0, 1, \dots, n$$

- Mean (p. 185) : $\mu = np$.
- Variance (p. 185) : $\sigma^2 = npq$.
- Tables : Contain $P(x \leq k)$ for $k = 0, 1, \dots, n$.



Chapter 5 : Continuous Random Variables

- Possible values are all those associated with one or more line intervals.
- Probabilities are areas under “density function”.



- If x is a continuous r.v.,

$$\begin{aligned} P(a \leq x \leq b) &= P(a < x \leq b) \\ &= P(a \leq x < b) = P(a < x < b) \end{aligned}$$

- Normal distribution is special case.

- Areas under normal curves between **z-scores** of 0 and z , for $z > 0$ in Table IV, p. 809.
- Key to finding correct areas (probabilities) : **draw pictures**

Chapter 6: If we plan to take a random sample of size n from a **population** with mean μ and standard deviation σ ,

- \bar{x} is a **random variable**.
- Dist. of \bar{x} is called its **sampling distribution**.
(p. 255)
- $\mu_{\bar{x}} = E(\bar{x}) = \mu$. So \bar{x} is an **unbiased** estimator of μ . (p. 266, 261)
- $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ (p. 266), so dist. of \bar{x} is more concentrated around μ for larger sample sizes. $\sigma_{\bar{x}}$ is called the **standard error** of \bar{x} . (p. 266)
- If the population has a normal dist., then so does \bar{x} , i.e., $\bar{x} \sim N(\mu, \sigma / \sqrt{n})$. True for any n .

- **Central Limit Theorem (CLT):** (p. 267) If n is large ($n \geq 30$), then the sampling distribution of \bar{x} is approximately normal, i.e., $\bar{x} \sim N(\mu, \sigma / \sqrt{n})$, regardless of the shape of the population distribution.

Chapter 6:

- A **Parameter** is a meaningful number associated with a Population. μ, σ^2 , etc. (p. 254)
- A **Statistic** is a meaningful number associated with a Sample.
- All statistics have **sampling distributions**

Chapter 7:

- Point Estimator (p. 261)
- Interval Estimator (p. 282)
- Confidence Coefficient (p. 282)
- Confidence Level (p. 282)

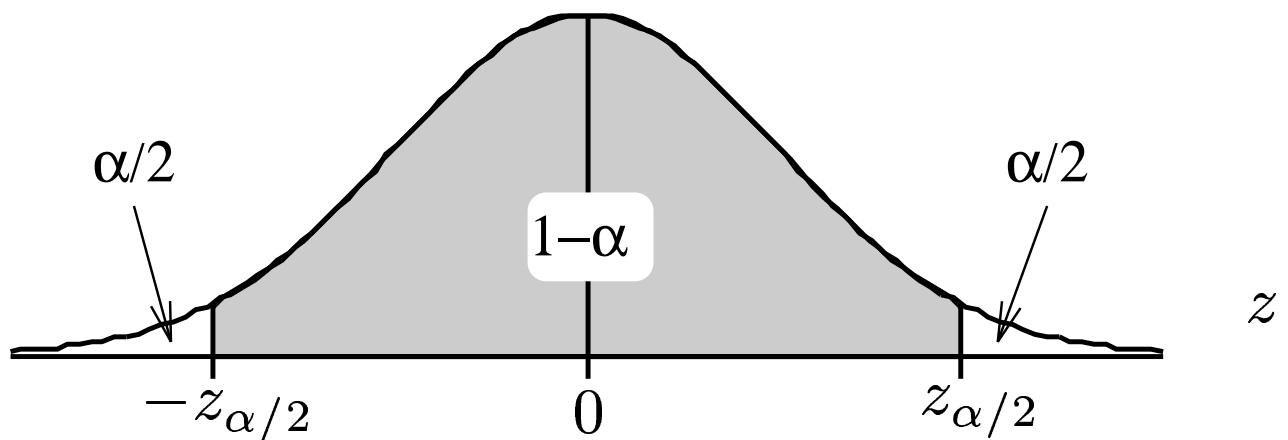
Parameter	Estimator	Standard Error of Est.	Estimated Standard Error of Est.
μ	\bar{x}	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
p	$\hat{p} = \frac{x}{n}$	$\sqrt{\frac{pq}{n}}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$

- Both estimators are UNBIASED
- If n is “large”, both estimators are approximately NORMALLY distributed.
- How large is “large”?
 - For valid CI for μ : $n \geq 30$.
 - For valid CI for p :

$$n \geq 9 \left(\frac{\text{larger of } (\hat{p}, \hat{q})}{\text{smaller of } (\hat{p}, \hat{q})} \right)$$

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

$$\underbrace{\text{estimator}}_{\text{formula sheet}} \pm \underbrace{z_{\alpha/2}}_{\text{table}} \underbrace{(\text{standard errors})}_{\text{formula sheet}}.$$



- Population mean, μ (P. 283)

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \approx \bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

- Population Proportion, p (P. 300)

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Finding the sample size to estimate μ .

- Want : Correct to within “ B ” units with $(1 - \alpha)100\%$ confidence.
- $z_{\alpha/2} \times (\text{standard error}) = B$ and SOLVE
- $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = B$ and solve for n (p. 307)
 - Use ballpark value for σ if you have one. Maybe use $\sigma \approx \frac{\text{Range}}{4}$.

Finding the sample size to estimate p .

- Want : Correct to within “ B ” units with $(1 - \alpha)100\%$ confidence.
- $z_{\alpha/2} \times (\text{standard error}) = B$ and SOLVE
- $z_{\alpha/2} \sqrt{\frac{pq}{n}} = B$ and solve for n (p. 333)
 - Use “ballpark” value for p if you have one, if not use $p = q = .5$ to get sample size that will work for **any** value of p .

Chapter 8 – Large Sample Hyp. Testing

Parts of a statistical test. (p. 322)

- The hypothesis of **MAIN INTEREST** is the **ALTERNATIVE** or **RESEARCH** hypothesis, – light bulb ex. H_a , ($H_a : \mu > 1325$) . (p. 322)
What we are “trying to prove” in an objective, fair manner
- The “**other**” hypothesis is called the **NULL HYPOTHESIS**, – H_o , light bulb ex. ($H_o : \mu = 1325$) (p. 322)

Errors: p. 325

<u>Decision</u>	Reality	
	H_o true	H_a true
Accept H_o	Correct	Type II error
Reject H_o	Type I error	Correct

- $\alpha = P(\text{Type I error})$ (p. 323), **SIGNIFICANCE LEVEL** of the test.
- $\beta = P(\text{Type II error})$ (p. 325)
- $\alpha \uparrow, \beta \downarrow$ and/or $\alpha \downarrow, \beta \uparrow$
- $\alpha = P\{\text{accepting } H_a \text{ when } H_o \text{ true}\}$
- $\alpha = P\{\text{saying what we "want" to say when we should not}\}$
- In our lightbulb example,
 $\alpha = P\{\text{saying } \mu > 1325 \text{ when } \mu = 1325\}$

Chapter 8 : Large Sample Tests about μ

- $H_o : \mu = \mu_0$: $z_{calc} =$ calculated value of z .

H_a : **RR** **p-value**

$$\mu > \mu_0 \quad z > z_\alpha \quad P(z > z_{calc})$$

OR

$$\mu < \mu_0 \quad z < -z_\alpha \quad P(z < z_{calc})$$

OR

$$\mu \neq \mu_0 \quad z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2} \quad 2 \times (\text{tail area})$$

- Test Statistic

$$z = \frac{\bar{x}^* - \mu_0^{**}}{\sigma / \sqrt{n}^*}$$

* Estimator and Standard Error from Formula Sheet

** Hypothesized Value from NULL HYPOTHESIS

p -value = **smallest** value for α for which H_0 can be rejected.

Large Sample Tests about p

- $H_o : p = p_0$: z_{calc} = calculated value of z .

H_a : **RR** **p-value**

$$p > p_0 \quad z > z_\alpha \quad P(z > z_{calc})$$

OR

$$p < p_0 \quad z < -z_\alpha \quad P(z < z_{calc})$$

OR

$$p \neq p_0 \quad z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2} \quad 2 \times (\text{tail area})$$

- Test Statistic

$$z = \frac{\hat{p}^* - p_0^{**}}{\sqrt{\frac{p_0 q_0}{n}}^{**}}$$

* Estimator and Standard Error from Formula Sheet

** Hypothesized Value from NULL HYPOTHESIS

Thought: *Eagles may soar, but weasels aren't sucked into jet engines.*

Assignments

Today : Minitab, Computer Demo,

P. 288 – 393 (Sec. 7.2), P. 341 – 345 (Sec. 8.4)

For Thursday: **Exer.** 7.27, 7.30, 7.33, 7.80, 7.81, 8.49,
8.50, 8.53, 8.54, 8.56, 8.57, 8.105 – 108, 8.111,
8.117

Summary: Large Sample Hypothesis Tests

- H_o : param. = value : $z_{calc} =$ calc. value of z .

H_a :	RR	p-value
param. $>$ value	$z > z_\alpha$	$P(z > z_{calc})$

OR

param. $<$ value	$z < -z_\alpha$	$P(z < z_{calc})$
------------------	-----------------	-------------------

OR

param. \neq value	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$2 \times$ (tail area)
---------------------	--	------------------------

- Test Statistic

$$z = \frac{\text{estimator}^* - \text{hypothesized value}^{**}}{\text{standard error}^*}$$

* Estimator and Standard Error from Formula Sheet

** Hypothesized Value from NULL HYPOTHESIS

Ex. : #8.68, p. 352 In a “Pepsi Challenge”, 100 Diet Coke drinkers were given unmarked cups of both Diet Coke and Diet Pepsi. 56 indicated that they preferred the taste of Diet Pepsi. Is there sufficient evidence to indicate that a majority of the Diet Coke drinkers will select Diet Pepsi in a blind taste test?

- p = true proportion of Diet Coke drinkers who would select Diet Pepsi in a blind taste test.

$$H_a : p > .50 \quad (1)$$

$$H_o : p = .50 \quad (2)$$

- $\alpha = .05$ level test, RR : $z > z_\alpha = z_{.05} = 1.645$
- Assumptions : the 100 individuals participating in the the Pepsi Challenge are a RANDOM SAMPLE of all Diet Coke drinkers. Note: n is “large”

- Data : $n = 100$ $\hat{p} = \frac{56}{100} = 0.56$

$$z = \frac{.56 - .50}{\sqrt{\frac{.50 \times .50}{100}}} = \frac{.06}{.05} = 1.20.$$

- Conclusion : is $z = 1.20 > 1.645$?

NO!! - Cannot reject H_o in favor of H_a **AT THE**
 $\alpha = .05$ **LEVEL!!**

- In terms of this problem:

“ CANNOT claim that there is sufficient evidence at the .05 level of significance” (or with 95% confidence) to indicate that the majority of Diet Coke drinkers will select Diet Pepsi in a blind taste test.

- p -value?

- p -value = $P(z > 1.20) = .5 - .3849 = .1151$.

Minitab?

- Stat→Basic Statistics→1 Proportion
- Click radio button “Summarized Data”, type in Number of trials, Number of Successes
- Click Options, Select Alternative, Type in Null Value
- Click Box “Use test and interval based on normal distribution”, OK, OK

Test and Confidence Interval for One Proportion

Test of $p = 0.5$ vs $p > 0.5$

Sample	X	N	Sample p	90% CI	Z-Value	P-Value
1	56	100	0.560000	(0.462710, 0.657290)	1.20	0.115

Computer Study:

- $H_o : p = .5$ $H_a : p \neq .5$
- $\alpha = .10$; RR : $z > 1.645$ or $z < -1.645$

$$z = \frac{\hat{p} - .5}{\sqrt{\frac{.5 \times .5}{n}}}$$

Sample size for each test is $n = 30$

p	reject H_o	not reject H_o	# tests	Prop. rejects
.5	4	46	50	.08
.6	12	38	50	.24
.7	32	18	50	.64
.8	48	2	50	.96
.2	47	3	50	.94
.1	50	0	50	1.00

Sample size for each test is $n = 60$

p	reject H_o	not reject H_o	# tests	Prop. rejects
.5	6	44	50	.12
.6	21	29	50	.42
.7	48	2	50	.96
.8	50	0	50	1.00

What do we see?

- For each fixed sample size, as the value of p moves **away from** .5, (and the null becomes “less true”) we REJECT H_o a **greater percentage** of the time.
— GOOD!
- When $p = .5$, for each n we reject H_o approx. 10% of the time. ($\alpha = .10$).
- For each fixed value of $p \neq .5$, we REJECT H_o a **greater percentage** of the time for **larger** n . Big n is “better”.

Small Sample Inferences about μ

$n < 30 \Rightarrow$ can't use CLT to get NORMALITY
of the sampling distribution of \bar{x}

\Rightarrow **can't use z -scores**

$\frac{\bar{x} - \mu}{s/\sqrt{n}}$ does **not** have a standard normal dist.

However:

- If the POPULATION is approximately NORMALLY distributed

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad (\text{looks a lot like } z!!!)$$

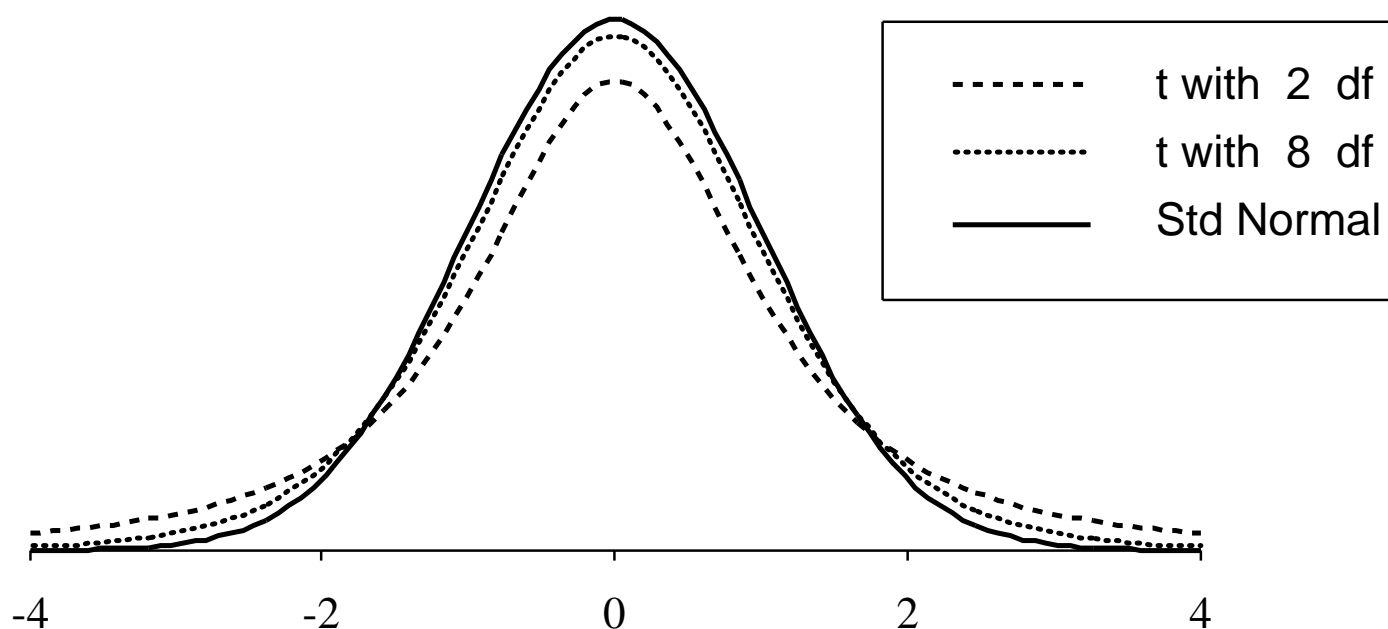
has a sampling distribution called the

t -**distribution** with $n - 1$ “degrees of freedom”,

d. f.

Properties of the t -distribution:

- Symmetric about 0. (like the z -distribution)
- Bell-shaped . (like the z -distribution)
- More variable (heavy-tailed) than the z -distribution
 - Variability depends on ***degrees of freedom***.
 - Variability \downarrow as d.f. \uparrow .
 - Becomes more and more like the z -distribution as d.f. \uparrow .



- **Note:** d.f. = denominator in calculating s^2 :

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

- Thus

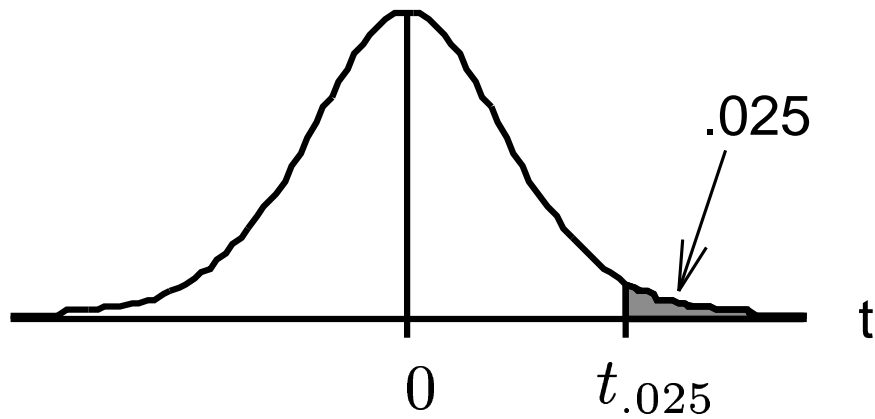
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

has the **same number of d.f.** at the estimator for σ used in its calculation.

- Define t_α so that $P(t > t_\alpha) = \alpha$
(Remember: z_α so that $P(z > z_\alpha) = \alpha$)
- Table VI (p. 811) gives t_α -values for $\alpha = .10, .05, .025, .01, .005, .001$ and $.0005$.

d.f.	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

d.f.	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576



$$\text{df}=5 \quad t_{.025} =$$

$$\text{df}=10 \quad t_{.025} =$$

$$\text{df}=20 \quad t_{.025} = 2.086$$

$$\text{df}=30 \quad t_{.025} = 2.042$$

$$\text{df}=\infty \quad t_{.025} =$$

Note : When d.f. = ∞ , $t_{\alpha} = z_{\alpha}$

Small Sample Inferences About μ

Assumption : POPULATION approx. NORMALLY dist.
Small sample situation similar to large, except use t dist. with $n - 1$ d.f. instead of z dist.

- Confidence Interval :

Large Sample:
$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Small Sample (p. 292):
$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

- Hypothesis Tests (p. 342)

$$H_o : \mu = \mu_0$$

versus

$$H_a \left\{ \begin{array}{ll} \mu > \mu_0 & t > t_\alpha \\ \text{OR} & \\ \mu < \mu_0 & t < -t_\alpha \\ \text{OR} & \\ \mu \neq \mu_0 & t < -t_{\alpha/2} \text{ or} \\ & t > t_{\alpha/2} \end{array} \right\} \text{RR}$$

- Test statistic :

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (\text{looks just like } z!!)$$

- t_α (and $t_{\alpha/2}$) depends on α (like before) AND #d.f.
(new)

Ex. Phosphorus content is a water quality index that is of concern to the EPA in the Everglades. In one section of park, EPA makes $n = 12$ measurements, obtaining $\bar{x} = 12.3$ and $s = 5.4$ (measurements in parts per billion [ppb]). Can the EPA support the claim that the mean level of phosphorus is less than 15ppb? Use $\alpha = .05$.

- $H_a : \mu < 15$ $H_o : \mu = 15$

- Rejection Region: Lower tail test. $n =$,
d.f. = , $t_\alpha = t_{.05} =$.
reject H_o if t

- Test statistic: $\bar{x} = 12.3$ $s = 5.4$

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{12.3 - 15}{5.4/\sqrt{12}} = -1.732$$

- Conclusion: Since $t = -1.732$ is in the rejection region, CANNOT reject H_o . There is evidence to conclude, at the $\alpha = .05$ level of significance, that the mean level of phosphorus is less than 15ppb.

Ex. Give a 95% CI for the mean phosphorus index in the section of the Everglades

$$n = 12 \quad \bar{x} = 12.3 \quad s = 5.4$$

$$\alpha = \quad t_{\alpha/2} = t_{.025} =$$

95% CI is

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} =$$

$$=$$

Note: In last example (both test and CI), we are assuming that population from which the sample is taken is (approx) normally distributed

That is, that
are (approx) normally distributed