Statistics 101 Review for Final

1. Introduction and Sampling

   A. Use data to say something about a characteristic of the population.

   B. Data are viewed as a sample from a population

   - Each item is chosen in the same way.
   - The items are chosen independently (in an unrelated manner).

   C. Errors that can arise in sampling from a finite population

   - Sample from an incorrect population.
   - Non-response bias (people who do not respond are different from people who do respond).
   - Response bias (the responses are not honest).

   D. 1. In the first part of the course we estimated various quantities

   - $\overline{x}$ for μ
   - s for σ
   - $\widehat{p}$ for p
   - Other quantities such as sample correlation r for ρ

   2. Variability of estimates across samples

   - $Var\left(\overline{X}\right) = \sigma^2/n$
   - $Var\left(\widehat{p}\right) = p(1-p)/n$

   3. Now we incorporate the variability to draw inferences

2. Confidence Intervals

A. 1. Given α (creates a $100\,(1-\alpha)\,\%$ confidence interval) find the range A to B that includes the true population characteristic with probability $1-\alpha$.

2a. Often it is of the form: estimate $\pm$ margin of error

b. Often the margin of error is z* $\times$ standard error of estimate or t* $\times$ estimate of standard error of estimate

B. Special cases

- μ where σ is known: $\overline{x} \pm z^{*}\sigma/\sqrt{n}$
- μ where σ is unknown: $\overline{x} \pm t^{*}s/\sqrt{n}$
- p. Let $\widetilde{p} = \frac{\#successes+2}{n+4}$ $\widetilde{p} \pm z^{*}\frac{\sqrt{\widetilde{p}(1-\widetilde{p})}}{\sqrt{n+4}}$ or $\widehat{p} \pm z^{*}\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}$

Notes: 1. $z^{*}$is the z-value corresponding to $1-\alpha/2$ in the Z-Table

2. $t^{*}$is the entry in the $n-1^{st}$ row and $\alpha/2$ column in the t-Table

C. Sample Sizes

- μ where σ is known: $n = \frac{(z^{*})^{2}\sigma^{2}}{E^{2}}$produces an interval of $\pm E$
- p $n = \frac{(z^{*})^{2}1/4}{E^{2}}$ produces an interval of $\pm E$

Note: If we know p is bounded away from 1/2 (e.g., $p \leq .1$) replace 1/4 with $p(1-p)$ for the p in the feasible region that is closest to 1/2.

3. Introdcution to Hypothesis Testing

A. Choose H_0 and H_a What we want to show is H_a

B. Structure and Terminology

Hypothesis	Null H_0 is true	Alternative H_a is true
Retain H_0	Correct	Type II error $\beta = P(Type\,II\,error)$
Reject H_0	Type I error $\alpha = P(Type\,I\,error)$	Correct

C. We create tests for a specified level of α

P-value= Prob(observing the sample at hand or one more extreme when H_0 is true)

If P-value $< \alpha$ then we reject the null hypothesis

4. Hypothesis Tests for Population Characteristics

A. Basics

Hypothesis Testing

$Hypothesis$	$H_0 : \mu = \mu_0$ $Ha : \mu \neq \mu_0$	$H_0 : \mu \leq \mu_0$ $Ha : \mu > \mu_0$	$H_0 : \mu \geq \mu_0$ $Ha : \mu < \mu_0$
Acceptance Region	$\mu_0 \pm z^* \frac{\sigma}{\sqrt{n}}$	Up to $\mu_0 + z_1^* \frac{\sigma}{\sqrt{n}}$	From $\mu_0 - z_1^* \frac{\sigma}{\sqrt{n}}$
$Z - score$ $z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$If\ \lvert z \rvert \geq z^*$ $reject$	$If\ z \geq z_1^*$ $reject$	$If\ z \leq -z_1^*$ $reject$
$P - value$	$2P\left(Z \geq \lvert z \rvert\right)$	$P\left(Z \geq z\right)$	$P\left(Z \leq z\right)$
Proportions $z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}$	$If\ \lvert z \rvert \geq z^*$ $reject$	$If\ z \geq z_1^*$ $reject$	$If\ z \leq -z_1^*$ $reject$
T-score $t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$	$If\ \lvert t \rvert \geq t^*$ $reject$	$If\ t \geq t_1^*$ $reject$	$If\ t \leq -t_1^*$ $reject$

Notes

- z^* is the z-value associated with the probability of $1 - \frac{\alpha}{2}$
- z_1^* is the z-value associated with the probability of $1 - \alpha$

For proprotions:

- μ_0 is replaced by p_0
- σ is replaed by $\sqrt{p_0(1-p_0)}$

In the case of unkown variance:

- σ is replaced by s
- $z^*$is replaced by t^* where $t^*$is the value in the $\frac{\alpha}{2}$ column and

  $n - 1^{st}$ row of the t-table

- z_1^* is replaced by t_1^* where $t_1^*$is the value in the α column and

  $n - 1^{st}$ row of the t-table

B. Sample Size and Type II Error for μ with known σ^2

- Probabiliy of errors

a. Create range of $\overline{x}$ to retain the null hypothesis (i.e., acceptance regions)

i) $H_0 : \mu = \mu_0$ $\quad \mu_0 \pm z^* \frac{\sigma}{\sqrt{n}}$
ii) $H_0 : \mu \leq \mu_0$ $\quad$ up to $\mu_0 + z_1^* \frac{\sigma}{\sqrt{n}}$
iii) $H_0 : \mu \geq \mu_0$ $\quad$ From $\mu_0 - z_1^* \frac{\sigma}{\sqrt{n}}$

b. For any assumed value of μ referred to as μ_a
Find the probability that $\overline{X}$ is in the acceptance region.
Use the fact that $\overline{X}$ is Normal with mean=$\mu_a$and
standard deviation=$\frac{\sigma}{\sqrt{n}}$

c. If μ_a belongs to H_a then the probability in b. is a Type II error
If μ_a belongs to H_0 then the probability in b. is not making a Type I error

- Find the sample size to produce a Type II error of a given β at a given α.

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{(\mu_a - \mu_0)^2}$$

C. Related Tests

- Paired Data- Pairs of (x, y) values generally taken on the same individual or under a common condition

a. Take differences $d = x - y$ for each case.
Treat it as a one sample problem as above.

b. Call a success whenever $x > y$ and a failure whenever $x < y$
(eliminate observations when $x = y$ and reduce n accordingly).
Treat the problem as a test of proportions as above with $p_0 = 1/2$.

- Tests for Medians- Call each x that exceeds M_0 a success; call each x that is less than M_0 a failure

(eliminate observations for which $x = M_0$ and reduce n accordingly).
Treat the problem as a test of proportions with $p_0 = 1/2$.