Midterm 2

STAT 324

Midterm: Thursday, April 3rd

Confidence Intervals – Single Variable

General Equation…

$$point\ estimate \pm test\ statistic \cdot SE$$

Point estimate…

- Either sample mean or sample proportion

Test statistics…

z-test…

- Normal curve
- z-score
  - Must know population standard deviation σ

$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma}$	$z_p = \frac{p - \pi}{\sqrt{\pi(1-\pi)}}$

```
qnorm(conf_level+(1-conf_level)/2)
```

t-test…

- Normal curve

```
qt(conf_level+(1_conf_level)/2, df=n-1)
```

Standard error…

z-test

- Assume $\pi = p$
- Assume $\mu = \bar{X}$

$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$	$SE_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

t-test

- Don't know population standard deviation → use sample s_x

$SE_{\bar{X}} = \frac{s_{\bar{X}}}{\sqrt{n}}$	$SE_p = \sqrt{\frac{p(1-p)}{n}}$

Confidence Intervals – Two Variables

Difference of means…

$$\mu_A - \mu_O$$

Test statistics…

z-test…

```
z.crit=qnorm(conf_level+(1-conf_level)/2)
```

t-test…

```
t.crit=qt(conf_level+(1-conf_level)/2,df)
df = n1+n2-2 = (n1-1)+(n2-1)
```

Standard error…

z-test…

$SE_{\bar{X}} = \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$	$SE_p = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$

t-test…

$SE_{\bar{X}} = \frac{s_1}{\sqrt{n_1}} + \frac{s_2}{\sqrt{n_2}}$	$SE_p = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Hypothesis Testing – Single Variable

Steps…

(1) Prove assumptions
   a. Sample IID → independent draws from one another
   b. Prove normal distribution (must satisfy both equations)

$$n(1-p) \geq 10$$

$$np \geq 10$$

`qqnorm(data)`

(2) Create hypothesis
   - Only use population proportions μ, π in hypothesis statements
   a. H_O → what's assumed to be true
   b. H_A → what you're trying to prove/opposite of H_O

(3) Find test statistics
- Change SE equation depending on z/t

$\frac{\bar{X}-\mu}{SE}$	$\frac{p-\pi}{SE}$

(4) Find p value depending on z/t

`p.z=pnorm(z)`	`p.t=pt(t,df=n-1)`

(5) Conclusion
   a. Given α and p is the probability of a type I error
      i. If $p < \alpha$ → sufficient data to reject H_O
      ii. If $p > \alpha$ → insufficient data to reject H_O

Hypothesis Testing – Medians

ONLY SINGLE VARIABLE

Steps…

(1) Prove assumptions
   a. Sample IID → independent draws from one another
   b. Prove normal distribution (must satisfy both equations)

```
qqnorm(data)
```

(2) Create hypothesis
   - Only use median M in hypothesis statement
   a. H_O → what's assumed to be true
      i. $M_O = some\ number$
   b. H_A → what you're trying to prove/opposite of H_O

(3) Perform sign test

$$n = n_+ + n_-$$

$$n_- = number\ of\ minus = number\ of\ values\ below\ M_O$$

$$n_+ = number\ of\ positives = number\ of\ values\ above\ M_O$$

(4) Determining test statistic S
   a. If $M > M_O$ → $S = n_-$
   b. If $M < M_O$ → $S = n_+$
   - THE PROOF
      i. $M \neq M_O$ → $S = \min(n_+, n_-)$
      ii. Determine significance
         1. $n > 25$

$$z = \frac{S + 0.5 - n/2}{\sqrt{n}/2}$$

      iii. Find p value using z score

```
p=2*qnorm(z)
```

<u>Power</u>

Finding power from a hypothesis test → z-test only

(1) Make a rejection region

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \pm\left(\frac{z^*}{t^*}\right)$$

$$\bar{X} = \left(\pm(z^*/t^*)\frac{\sigma}{\sqrt{n}} + \mu\right)$$

(2) Find upper/lower crit values

a. $\bar{X}_{obs} = \mu_A$

$$z^*_{lower} = \frac{\bar{X}_{lower} - \bar{X}_{obs}}{\frac{\sigma}{\sqrt{n}}}$$

$$z^*_{upper} = \frac{\bar{X}_{upper} - \bar{X}_{obs}}{\frac{\sigma}{\sqrt{n}}}$$

(3) Find power

```
p.lower=pnorm(z.crit.lower)

p.upper=1-pnorm(z.crit.upper)

power=p.lower+p.upper
```

Determining the sample size n required to achieve a certain power

```
power.t.test(n=NULL,delta=xbar-mu,sd,signif_level=alpha,power,

type="[two.sample,one.sample,paired]",
alternative="[two.sided,one.sided]")
```

Hypothesis Testing – Two Variables

Steps…

(1) Prove assumptions
   b. Sample IID → independent draws from one another
   c. Prove normal distribution (must satisfy both equations)

$$n_1(1-p_1), n_2(1-p_2) \geq 10$$

$$n_1 p_1, n_2 p_2 \geq 10$$

```
qqnorm(data)
```

(2) Create hypothesis
   a. Differences of means $\mu_A - \mu_B$ or difference of proportions $p_A - p_B$
   a. H_O → what's assumed to be true
   b. H_A → what you're trying to prove/opposite of H_O

(3) Should it be pooled?
   a. Only if $Var(\bar{X}_1) \approx Var(\bar{X}_2)$

$$Sd(pooled) = s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

$$p_{pooled} = p_p = \frac{X_1 + X_2}{N_1 + N_2}$$

$$X_i = numerator\ of\ proportion\ i$$

$$N_i = deniominator\ of\ proprtion\ i$$

(4) Find standard error

$SE_{\bar{X},z} = \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$	$SE_{\bar{X},t} = \frac{s_1}{\sqrt{n_1}} + \frac{s_2}{\sqrt{n_2}}$	$SE_{\bar{X},pooled} = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
$SE_{p,z} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$	$SE_{p,t} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$	$SE_{p,pooled} = \sqrt{p_p(1-p_p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

(5) Find test statistics

(4) Change SE equation depending on $z/t/pooled$

$\frac{\bar{X}-\mu}{SE}$	$\frac{p-\pi}{SE}$

(6) Find p value depending on z/t

`p.z=pnorm(z)`	`p.t=pt(t,df=n-1)`

(7) Conclusion

a. Given α and p is the probability of a type I error

i. If $p < \alpha$ → sufficient data to reject H_O

ii. If $p > \alpha$ → insufficient data to reject H_O

Additional Tests & Error

Type I error

- Rejecting H_O, when H_O is true

$$\alpha = P(type\ I\ error)$$

$$CL = 1 - \alpha$$

Type II error

- Failing to reject H_O, when H_O is false

$$\beta = P(type\ II\ error)$$

$$power = 1 - \beta$$

- People shoot for $power \approx 0.8$

Welch's t-test

(5) When we assume nonequal variance → $Var(\bar{X}_1) \neq Var(\bar{X}_2)$
(6) Used when trying to compare $\bar{X}_1$ and $\bar{X}_2$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

Bootstrap

(7) Process of analysis
   a. You will have a t_{obs}
   b. Run the bootstrap function, record

```
t.lower=t.i<=t.obs

p.lower=(t.i<=t.obs)/n.boot=t.lower/n.boot

t.upper=t.i>=t.obs

p.upper=(t.i>=t.obs)/n.boot=t.upper/n.boot
```

   c. ONE-SIDED: if you're trying to prove one, use the other
   d. TWO-SIDED: Find $2 \cdot \min(t_{lower}, t_{upper})$

Wilcoxon Rank-Sum Test

(8) Data is *not* normal and violates test assumptions
(9) Options for data…
   a. Ranked observations → observations ranked from least to greatest, or null
   b. Continuous data
(10) Small sample size
(11) Instead of comparing means, compares rankings of data sets to see if the rankings are different from one another
(12) Steps
(1) Declare two groups (given)
(2) Make hypotheses
(3) Combine two data sets
(4) Sum the ranks of the contents of each individual data set based on global rank

   Ex) take $A = [1,11,3]$ and $B = [5,10,13]$:

   $A + B = [1,3,5,10,11,13]$. $R_A = 1 + 5 + 2, R_B = 3 + 4 + 6$

(5) Find U_1, U_2 statistics

$$U_i = R_i - \frac{n_i(n_i + 1)}{2}$$

(6) Find z^*

$$z^* = \frac{\min(U_1, U_2) - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}}$$

(7) Find probability

```
pnorm(z.crit)
```

(8) Make conclusion