Sampling distribution

Conditional Distribution

$f_{Y|X} = \frac{f_{XY}(x,y)}{f_X(x)}$

If independent:

$f_{XY}(x,y) = f_X(x) f_Y(y)$

so $f_{Y|X}(y|x) = f_Y(y)$

Law of Large Number

For LLN means

Sample $(x_1, x_2, \dots x_n)$

$\bar{x} = \frac{1}{n}\sum_i x_i$ is sample mean

as $n \to \infty$, $\bar{x} \to E(x)$

LLN mediums

m is sample mean

As $n \to \infty$, $m \to M$

Central limit Theorem

Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ (mean of first n obs.)

$Var(x) = \sigma^2$

Then: $\sqrt{n}(\bar{x}_n - E(x)) \sim N(0, \sigma^2)$

Let $S_n = \sum_{i=1}^{n} x_i$, Then

$\frac{S_n - nE(x)}{\sqrt{n}} \sim N(0, \sigma^2)$

SD is more often used to refer to the individual observations whereas SE is the SD of the sampling distribution

Neyman-Pearson Lemma

Of all tests of Ho vs. Ha with given sig. level α, the likelihood ratio test has the largest power

$P(T > C \mid H_a)$

Likelihood ratio

$\frac{L(\theta_0 \mid \dots)}{L(\theta_a \mid \dots)}$

Type I error (Drug example)

Drug is claimed effective when it's not

Type II error (Drug example)

Drug is claimed ineffective when it is

$P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 = T)$

$\beta = P(\text{Type II error}) = P(\text{Don't reject } H_0 \mid H_A = T)$

$1 - \beta = 1 - P(\text{Reject } H_0 \mid H_A = T)$ (power)

Rejection Region:

Region in which the probability of the null hypothesis is true is lower than the defined significance level

Permutation test

comparing means

observed mean comparison

used sampling to see varying differences

checked to see how many sample difference are greater than observations

- that percentage is the p-value

goodness of fit

Apply chi-square to a bunch of samples, then follow same logic as above

two-way tables

one sample isn't change

Other sample's order is randomized

-check difference in mean

follow remain steps

Bootstrapping

1) Draw a resample of size n (size of pop.) with replacement

Compute statistic (mean/median/etc.)

2) Repeat at least 10,000 times

3) Map dist. of Dist

-To find SE of bootstrap, compute sd of Distribution

-bias is mean of Dist. - mean(pop)

two-sample

sample each sample independent of the other

small sample size with mean that the mean varies

Confidence Interval for mean

-If σ known, then use z score

-If σ is unknown, then calculate sample SD (S) and use t-score

When looking at Diff. of means

$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

use t-statistic if σ_1, σ_2 unknown

$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$

with df:

$\frac{(S_1^2/n_1 + S_2^2/n_2)}{(S_1^2/n_1)^2/(n_1 - 1) + (S_2^2/n_2)^2/(n_2 - 1)}$

T-Confidence Interval

$(\bar{X} - \bar{Y}) \pm q\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

when q is T-score at $(1 - \alpha/2)$ with above df

One-sided CI for mean

q is $(1-\alpha)$ quantile of t-distribution n-1 df

$\bar{X} - q \cdot \frac{S}{\sqrt{n}}$ (lower bound)

upper bound ∞

so ~~lower~~ upper confidence bound

$\left[\bar{X} - q \cdot \frac{S}{\sqrt{n}}, \infty\right]$

lower confidence band

$\left(-\infty, \bar{X} + q\, S/\sqrt{n}\right]$

Confidence Interval for prop.

use z-score for q: the $(1 - \alpha/2)$ quantile

$CI = \frac{\hat{p} + q^2(2n) - q\sqrt{\hat{p}(1-\hat{p})/n + q^2/4n^2}}{1 + q^2/n}$

(use prop.test in R)

CI for Difference in prop.

$\tilde{p}_1 - \tilde{p}_2 \pm q\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{\tilde{n}_1} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{\tilde{n}_2}}$

$\tilde{p}_i = (x_i + 1)/(n_i + 2)$

$\tilde{n}_i = n_i + 2$

bootstrap + Confidence Int.

-bootstrap method

-quantile for CI

Correlation coefficient

$\rho(x,y) = Cov(x,y)/(\sigma_x \sigma_y)$

Simple Linear Regression

$Y = \alpha + \beta x$

$\beta = r\frac{S_y}{S_x}$

$\alpha = \bar{Y} + \beta\bar{X}$

residuals: $r_i = y_i - (\alpha + \beta x_i)$

fitted Values

$\hat{y} = \alpha + bx$, $\hat{y}$ is the fitted Value aka predicted value.

Crib Sheet for Analytics 511 Final

12/14/15