Lecture 3

Wednesday, January 19, 2022 8:12 AM

Admin: ① No office hour today unless needed.
② VOP
③ HW Set 1 posted.

Objectives:

① Hypothesis testing
② Student's t-test (Normal test) as a modeling problem
③ permutation hypothesis testing.
④ Start on ordinary least squares regression.

Last time

1. properties of random variables

2. Hypothesis Testing

- A formalized approach for asking if there are "significant" differences between groups of data
  - → often used to ask if means are different between two groups.

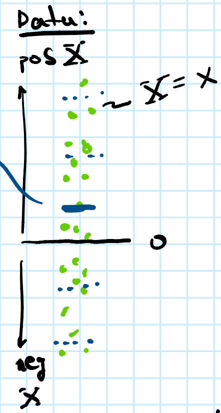
- Examples:
  1) is $E[X_1] \neq 0$?
  2) is $E[X_1] \neq E[X_2]$

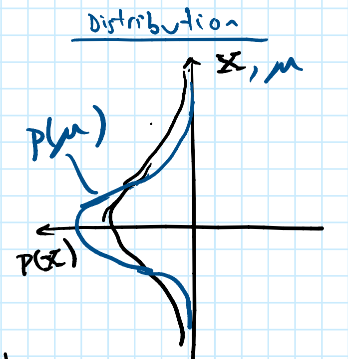
Formulate hypothes:

1. H_0 : $\mu = E[X] = 0$ (null hypothesis)
   H_1 : $\mu \neq E[X] \neq 0$ (alternative hypothesis)
2. Significance $\alpha = 0.05$.

want to know about the mean.

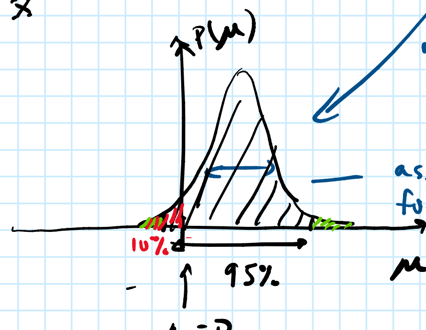
think mean is here





Estimate: μ, σ

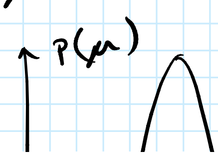
rotating 90°



assuming model for data.

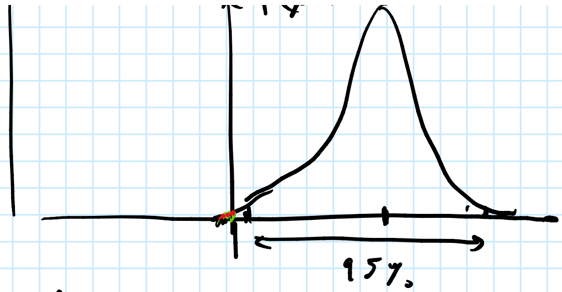
[Student's T-test] as a parameter estimation problem

Caveate bias in estimate of standard deviation due to

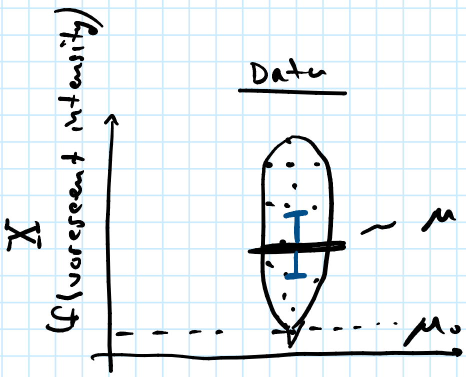


Caveate bias in estimate of standard deviation due to small sample size

we will ignore



- t-test / normal test : one type of hypothesis test.



- Null hypothesis : $\mu = \mu_0 \rightarrow P(\mu = \mu_0)$
- Alternative : $\mu \neq \mu_0$
- t-test/normal test: assumes a model for the data:

$$p(X) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

true std. dev.

true mean

- Need $\underline{\underline{P(\mu)}}$

- Estimation of mean:

$$\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i \Rightarrow E[\bar{\mu}] = \frac{1}{N}\sum_{i=1}^{N} E[X_i]$$

$= E[X]$

$$= \frac{1}{N}\{N E[X]\}$$

$$= E[X]$$

- we need $\sigma_{\bar{\mu}}$

- Recall that $Var[X] = E[X - E[X]]^2$

$$\Rightarrow Var[\bar{\mu}] = E[\bar{\mu} - E[\bar{\mu}]]^2$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N} X_i - E\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right]\right]^2$$

$\sum_i \underbrace{E[X_i]}_{E[X]}$

$$= \frac{1}{N^2} E\left[\sum_i X_i - N E[X]\right]^2$$

$$= \frac{1}{N^2} E\left\{ \underbrace{[X_1 - E[X]]} + \underbrace{[X_2 - E[X]]} + \cdots \right\}^2$$

indep. $\Rightarrow Cov(\ ) = 0$

indep. $\Rightarrow Cov(\ \ ) = 0$

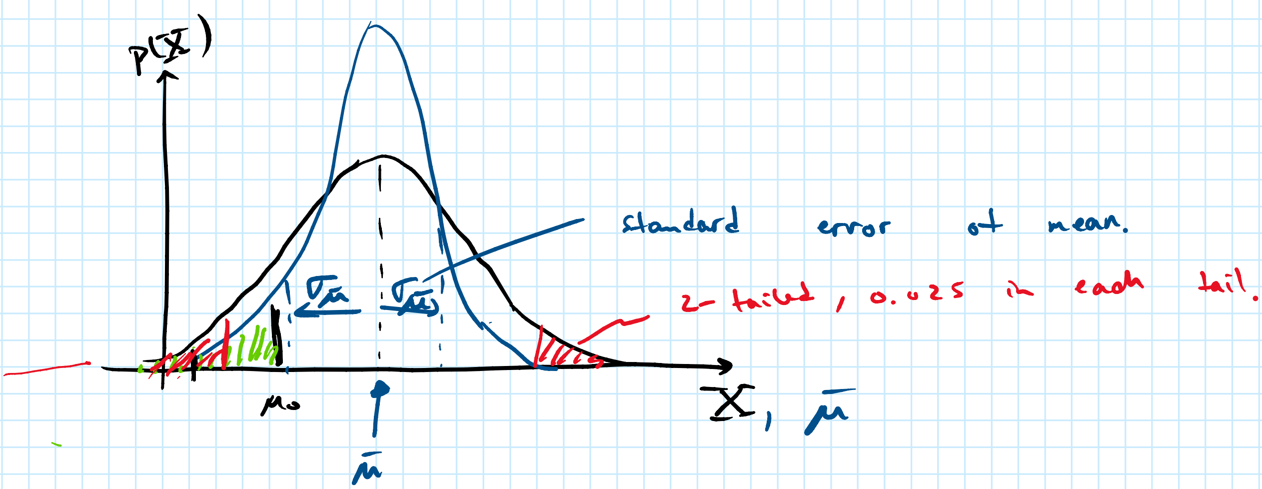
$$= \frac{1}{N^2}\left\{ \underbrace{E\left[X_1 - E[X_1]\right]^2}_{Var(X_1)} + \underbrace{E\left[X_2 - E[X_2]\right]^2}_{Var(X_2)} + \cdots \right\}$$

- But $var(X_1) = var(X_2) = var(X)$

$$\Rightarrow var[\bar{\mu}] = \frac{1}{N^2}\left\{ N\, var(X) \right\}$$

$$Var(\bar{\mu}) = \frac{1}{N}\, var(X)$$

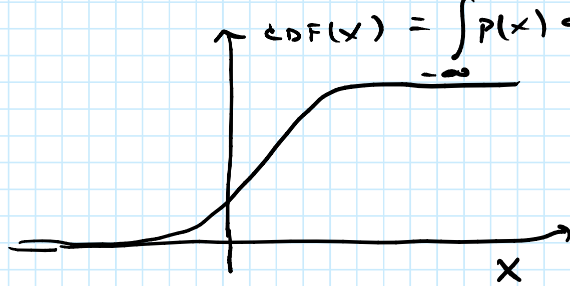
$$\Rightarrow \sigma_{\bar{\mu}} = \frac{1}{\sqrt{N}}\sigma_x \quad \Rightarrow p(\bar{\mu}) = \frac{1}{\sqrt{2\pi}\,\sigma_{\bar{\mu}}} e^{-\frac{1}{2}\left(\frac{\bar{\mu}-\mu}{\sigma_{\bar{\mu}}}\right)}$$



- Test null hypothesis $p(\mu = \mu_0)$

$\Rightarrow$ compute $Pr(\bar{\mu} \le \mu_0)$ $\Rightarrow$ cumulative density funcn in MATLAB or R.

$$CDF(x) = \int_{-\infty}^{x} p(x)\, dx$$



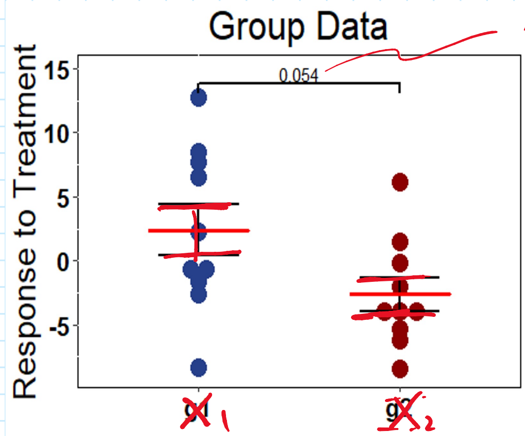
- T-test (Normal) test consists of

1. Estimation of mean
2. Estimation of variance of mean
3. Compute probability of null hypothesis.

$\Rightarrow p(\text{null hypotheses}) < 0.05$ or α (or 0.025 for two-tailed test)

$\Rightarrow$ assume alternative hypothesis

Consider permutation test to establish significant difference between two groups of data.



two - tailed t-test

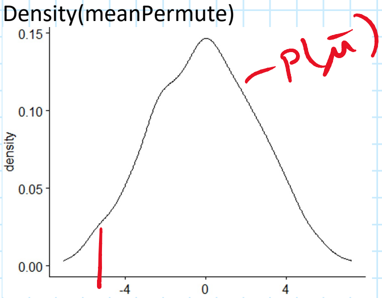
How to code a permutation test: goal is to compare the true difference in means with that derived from a random permutation of the data

meanTrue=mean(X2)-mean(X1) $\approx 5.$

x = c(x1, x2)

```
for (i in 1: no of permutations)  #1000 is a good number
{
    meanPermute[i]=mean(X[randInd2])-mean(X[randInd1])
}
```

Density(meanPermute)



$-p(\bar{\mu})$

$\bar{\mu}$

mean true

prob (Ho)

pPermute=mean(abs(meanPermute) > abs(meanTrue)) #Two-tailed test. One tailed would remove the absolute values

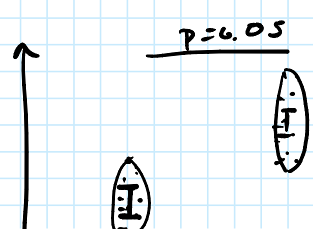
# of times that difference in permutation > true.

- most often extended to check for "significant" differences between groups.

- Analysis of variances for each variable

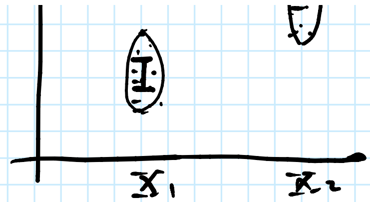
→ HW problem 1

→ $Z_1 = X_2 - X_1$



→ HW problem 1

→ $Z_1 = X_2 - X_1$



- These are <u>parametric</u> approaches
  - ↳ we have a model for the data.

- If model is unknown, can use alternative tests.

  - Kolmogorov - Smirnov test.
  - wilcox rank sum.
  - permutation test.

  } empirical distribution given data