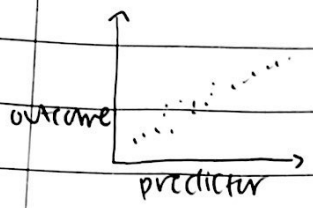
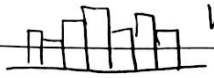
Week 3 Lecture Notes

Exploratory Data Analysis (EDA)



scatterplot


boxplot = compare 2 groups (categorical)


histogram (distribution)

Basics of Probability Distribution

Bernoulli = single trial w/ 2 outcomes P(success) P(failure) 1-P(success)

Binomial = multiple bernoulli samples strung together

Geometric = # of failures before success when strung together bernoulli trials

normal Dist = 

Poisson = (# of events over a time period (exp.)

Overview of Hypothesis Testing

① come up w/ research question ③ collect samples / data ⑥ Run your stats test

② Turn into hypothesis:
H_0 = no statistical difference
H_a = yes

④ EDA

⑤ decide signif. level

⑦ examine output/ make sure seems reasonable

⑧ Report your key stats / results

Statistical Tests (pt I) - z + t-tests

Population σ?

yes → n > 30 → yes: z-test; no: t-test

no → t-test

H_0 → 2 means not statistically different

H_A → yes difference

- one-sample z / t test = you are comparing a sample's mean to a predetermined mean (hypothesized mean)
- paired = comparing means of 2 samples from the same pop.
- 2sample = comparing means of a 2 samples, → 2 indp. samaples from 2 diff pop

Same indv before after treatment

indp: linked

σ ✓ + n ≥ 30 = z-test

σ x + n < 30 = t-test

t-test (assumption)

one-sample: 1) continuous
2) data follow a normal dist
3) Random sampling

paired: 1) continuous vars.
2) difference btwn the matched sample pairs is normaly dist
3) Random sampling.

2sample:
1) continuous data.
2) normaly distributed
3) samples r independent
4) variances of 2 samp are the same
5) random samp

z test assumptions

1-sample: 1) continuous data
2) data ≈ normal dist
3) sample is taken randomly
*4) pop SD KNOWN

paired z: 1) continuous data
2) differe b/w the pairs of data points follows a normal dist.
3) random sampling
*4) pop SD of the difference is known.

2sample 1) con. var.
2) random sample in both pops
3) independence of samples.
4) variance of dis of both pop. known

one tail: know if u is either <, > sample means
2-tail: difference (not directionality)

typical outputs:
1) t- or z-stat 2) df (n-1) 3) p value, confidence intv

ANOVA TEST

t-test, z-test → 2 samples
ANOVA → 3 or more groups
① one way ② two way
↳ 1 variable differs
↳ 2 categorical factors differ and how they effect each other

one way assumptions
1) normality (sample from aprox normal)
2) indp. samples.
3) variance across group same.
4) outcom is continuous
H0 = no difference, means equal
H1 = means differ

two-way assumptions.
1) outcome variable continous indepent.
2) group/vars should be categorical + independent
3) sample are independ.
4) variance accross groups the same
5) normality

*difference: # of groups

p-value & confidence interval

significance level (α) = 0.05
pvalue = prob that your result is due to random chance.

---

p < 5% (reject null)
p ≥ 5% (fail reject null)

that the pop mean is w/in
"95% of the CI will cover the true pop mean"

CI = (1-α) → α = 0.05 → 95% CI
provides a range of means = (1.02, 3.5)

↳ df (n-1) → critical value
↳ interpret test stat