Wed: HW + Lab activity
OH today 11:30-12:30, Wed afternoon
Stats talks: today + Thurs + next Monday

Quiz - back tomorrow at latest

10: HYPOTHESIS TESTING

Larsen & Marx 6.1-6.4

Prof Amanda Luby

So far, we've treated *inference* as either *point estimation* or *interval estimation*. In some experimental settings, however, we don't want to draw a numerical conclusion but rather evaluate two competing theories. For instance, we may wish to know whether a candidate for political office is likely to win or lose; whether a new vaccine is effective or ineffective; or whether a policy intervention improves or does not improve quality of life for citizens.

The process of dichotomizing possible conclusions from an experiment and using probability theory to choose one over the other is called *hypothesis testing*. We have two competing hypotheses:

1. H_0: null hypothesis
2. H_1: alternative hypothesis
   H_a

$H_0: \theta \in \Omega_0$
$H_1: \theta \in \Omega_1$

Parameter Space: all possible values of θ (where θ is some parameter)
Ω
$\Omega_0: \{\theta : \theta \in H_0\}$
$\Omega_1: \{\theta : \theta \in H_1\}$

Ω_1 and $\Omega_0 \in \Omega$
$\Omega_1 \cap \Omega_0 = \emptyset$

$\Omega_1 \cup \Omega_0$ is not necessarily $= \Omega$

Simple and Composite Hypotheses
simple: Ω_i contains a single θ
composite: Ω_i contains more than 1 θ

Typical set up:
H_0 is a simple hypothesis
H_1 is a composite hypothesis

1 Decision Rule

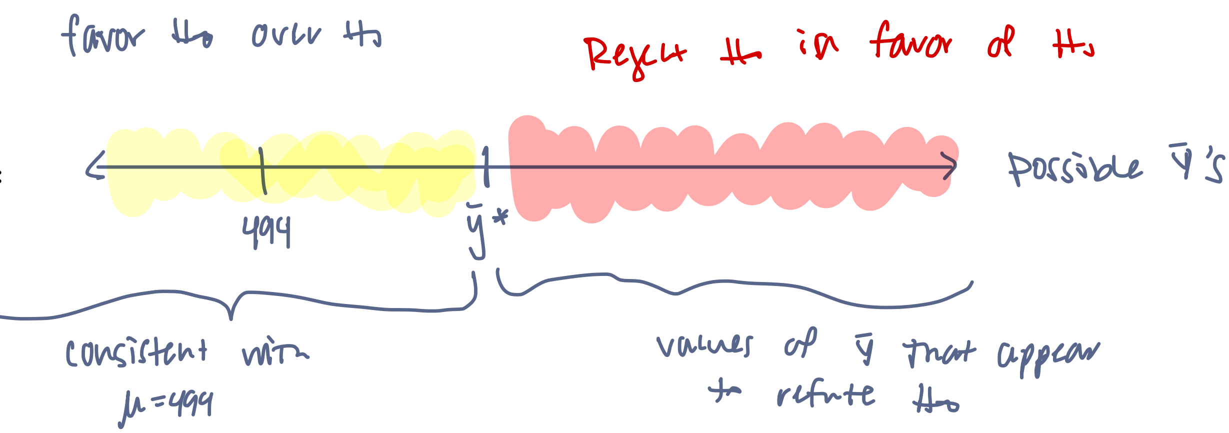
Example: A high school was chosen to participate in the evaluation of a new geometry and algebra curriculum. In the recent past, the school's students were considered "typical", receiving scores on standardized tests that were very close to the nationwide average. In the year of the study, 86 sophomores were randomly selected to participate in a special set of classes that integrated geometry and algebra. Those students averaged 502 on the SAT-I math exam; the nationwide average was 494 with a standard deviation of 124. Did the curriculum improve scores?

Assume: $Y_i \sim N(\mu, 124^2)$ where Y_i = student scores in class

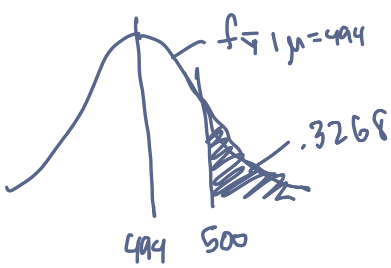
$H_0: \mu = \mu_0 = 494$
$H_1: \mu > 494$

Know $\bar{y} > 494$ (pop. mean) - is it bigger because scores in new curriculum are actually higher or is it due to random variation?

Decision Rule:



What if?



$\bar{y}^* = 500$ | $P(\text{reject } H_0 \mid H_0 \text{ true})$

$= P(\bar{y} \geq 500 \mid \mu = 494)$

$= P\left(\frac{\bar{y} - 494}{124/\sqrt{86}} \geq \frac{500-494}{124/\sqrt{86}}\right)$

$= P(Z \geq 0.44)$

$= .3268$

→ $\bar{y}^*$ is too small. Not strong enough evidence against H_0 to refute it

$\bar{y}^* = 600$

$P(\text{reject } H_0 \mid H_0 \text{ true})$

$\vdots$

$= 1.11 \times 10^{-15}$

→ $\bar{y}^*$ is too big - requiring too strong of evidence to refute H_0

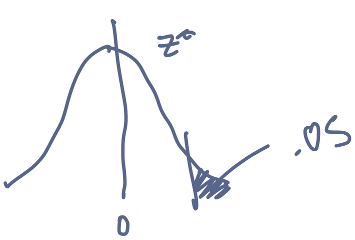
Mapping to Z-scores:

Rejecting when $\bar{y} \geq \bar{y}^*$ is equivalent to rejecting when $\frac{\bar{y} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{y}^* - \mu}{\sigma/\sqrt{n}} = z^*$

→ can use z-scores to decide on decision rule

Setting a significance level: Significance level $= \alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$

Can find a z^* value / decision rule that results in a α-level test



$z^* = 1.64$

If we want an $\alpha = .05$ test, define decision rule as: reject H_0 if $Z = \frac{\bar{y} - 494}{124/\sqrt{86}} > 1.64$

Test statistic

Numerical value that dictates when H_0 is rejected

$\bar{Y}$ and $z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ are both test statistics

Critical Region

Set of values that results in H_0 being rejected

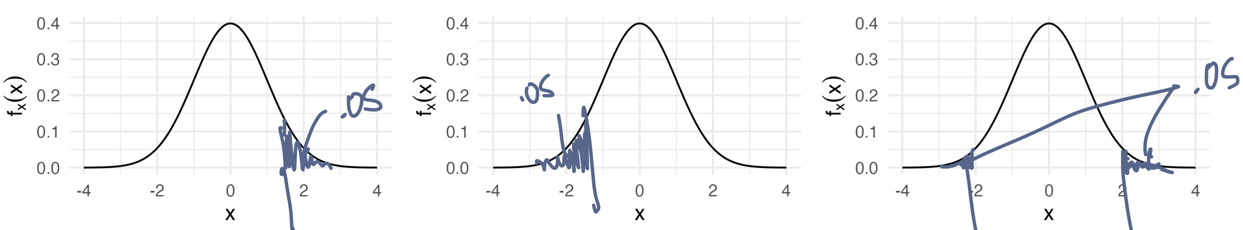
Sometimes denoted as C

Critical Value

Particular point in C that separates the "rejection" region from "acceptance" region

2 One-sided vs Two-sided Alternatives at $\alpha = .05$

$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$ | $H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$ | $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$



$z_{.05} = 1.64$ | $-z_{.05} = -1.64$ | $-z_{.025} = -1.96$ $z_{.025} = 1.96$

3 P-Values

There are two ways to quantify the amount of evidence against H_0 in a given dataset. The first involves defining a *level of significance* (α), identifying a corresponding *critical region*, and reject H_0 if the *test* *statistic* falls within the critical region. Another strategy is to calculate the p-value:

P-value Probability of observing a test statistic that is as or more extreme than what we actually observed if the null hypothesis is true

Example:

p-value: $P(\bar{Y} \geq 502 \mid \mu = 494) = P\left(\frac{\bar{Y} - 494}{124/\sqrt{86}} \geq \frac{502-494}{124/\sqrt{86}}\right)$

$= P(Z \geq .598)$

$= 0.27$ ← if H_0 is true, there's a .27 probability of observing $\bar{Y} \geq 502$.

reject H_0 if $\alpha > .27$

fail to reject when $\alpha < .27$

4 Non-normal data

Up to this point, we've assumed that we're working with the normal distribution and setting up a hypothesis test for a mean. Decision rules for other probability distributions are rooted in the same basic principles.

In general, to test $H_0 : \theta = \theta_0$, where θ is an unknown parameter in $f_x(x;\theta)$, we define the decision rule in terms of $\hat{\theta}$, a sufficient statistic for θ. We want to set up the decision rule such that the probability of rejection if the null hypothesis is true is equal to α.

Example: Four measurements $(k_1, ..., k_4)$ are taken of a Poisson random variable X (so $p_x(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$ and we wish to test $H_0 : \lambda = 0.8$ against $H_1 : \lambda > 0.8$.

$\alpha = .1$

$H_0 : \lambda = 0.8$

$H_1 : \lambda > 0.8$

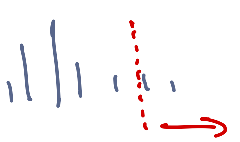
Recall: $\bar{X}$ is sufficient for λ

Dist of $\bar{X}$?

But we do know $\sum X_i \sim Pois(n\lambda)$

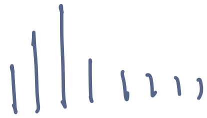
and $\sum X_i = f(\bar{X}) \Rightarrow \sum X_i$ is also sufficient

$\hat{\theta} = \sum_{i=1}^{4} X_i \sim Pois(3.2)$ under H_0.



$\approx .1$

Idea: want to find rejection region based on $\hat{\theta}$ where $P(\text{reject } H_0 \mid \lambda = 3.2) \approx 0.1 = \alpha$



K	P(X = K)
0	0.041
1	0.130
2	0.209
3	0.223
4	0.178
5	0.114
6	0.061
7	0.028
8	0.011
9	0.004
10	0.001
11	0.000
12	0.000
13	0.000
14	0.000
15	0.000

Probability = .105

PMF of a Pois(3.2) RV

Decision rule: reject H_0 when $\hat{\theta} > \theta^*$

$\hat{\theta} > 5$ → results in a $\alpha = .105$ level test.

If test statistic is discrete, we can't "set" α = any value, so often try to get "close enough"