Two-sample tests for $\mu_X - \mu_Y$ for non-normal populations

$H_0: \mu_X - \mu_Y = \Delta$

Set up: Same as before but the populations are non-normal

Test statistic:

$$z = \frac{\bar{x} - \bar{y} - \Delta}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \approx N(0,1) \text{ if } H_0 \text{ is true and both } n \text{ and } m \text{ are large.}$$

- Large-sample z-test. Its level is approximately α

Two-sample test for difference in proportions, $p_X - p_Y$

$H_0: p_X - p_Y = \Delta$

As before, apply large-sample z-test because the proportions can be interpreted as means of Bernoulli populations. Can also use *pooled sample proportion* in case of $H_0 : p_X = p_Y$ as suggested by the book.

Test statistic:

$$z = \frac{\hat{p}_X - \hat{p}_Y - \Delta}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}}$$

$\sim N(0,1)$ if H_0 is true & n and m are large.

- The level of the test is approximately α

Duality b/w testing and confidence intervals (two-sided case)

Suppose (L, U) is a $100(1-\alpha)\%$ confidence interval for θ.
(set of all plausible values of θ based on the data)

Q. Can we use this interval to do a level α test of

$$H_0: \theta = \theta_0 \quad \text{vs} \quad H_1: \theta \neq \theta_0.$$

A. Yes.

How?

Accept H_0 if $\theta_0 \in (L, U)$, otherwise reject H_0.
$\rightarrow$ level α test. [Already know: two other methods for doing level α tests].

- Another interpretation of CI: [Getting a CI from a test] $\rightarrow$ level α

$(L, U) = \{$All θ_0 for which $H_0: \theta = \theta_0$ is accepted$\}$.

- $100(1-\alpha)\%$ CI for θ.
- Connection holds for exact tests.

Duality in one-sided case

Case 1: $H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$

Suppose have a level-α test. Then, we can see that

$100(1-\alpha)\%$ CI for θ = {All θ_0 for which $H_0: \theta = \theta_0$ is accepted}

$= (L, \infty)$

L: lower confidence bound (smallest plausible value based on data)

Case 2: $H_0: \theta_0 = \theta_0$ vs $H_1: \theta < \theta_0$

Suppose have a level-α test. Then, we can see that

$100(1-\alpha)\%$ CI for θ = {All θ_0 for which $H_0: \theta = \theta_0$ is accepted}

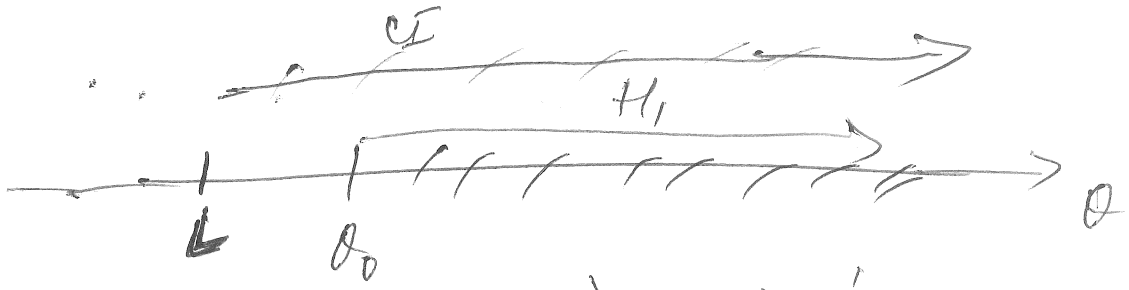
$= (-\infty, U)$

U: upper confidence bound (largest plausible value based on data)

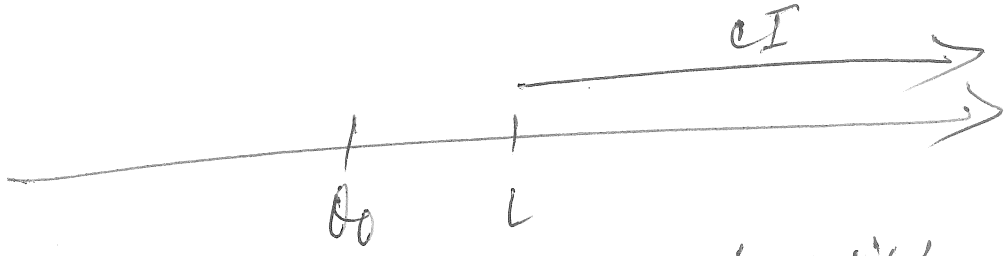
Suppose have a $100(1-\alpha)\%$ lower conf. bound for θ

— (L, ∞)

Q. How can we do a test of $H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$?



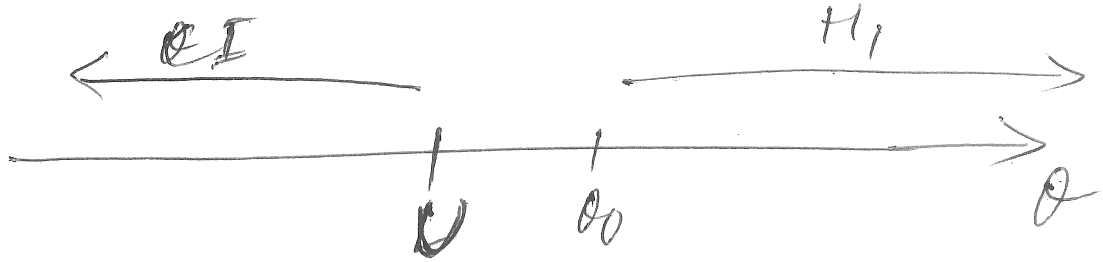
Accept H_0 since θ_0 is 'plausible'
falls in the CI.



Reject H_0 since θ_0 is not plausible

$H_0: \theta = \theta_0$ vs $H_1: \theta > \theta_0$

U = largest plausible value for θ based on data.



Regression (Chapter 11)

Setup: Have data on two *quantitative* variables — X and Y — on a sample of n subjects.

Q: Is there any association between X and Y? What kind?

Scatterplot:

- Plot y against x
- Look for the **trend** in the plot — a smooth curve that shows how the average value of Y changes with x
- Trend may be linear or non-linear
- If there is a trend, then the two variables are associated. In this case, x may be used to predict y
- Trend may be strong or weak. It is strong if the points are tightly clustered around the trend (small scatter)
- No trend: No association — i.e., the variables are independent, and x is not helpful for predicting y.

Example: House price data

```
house <- read.table(file="house_price.txt", sep=",",
header=T)
```

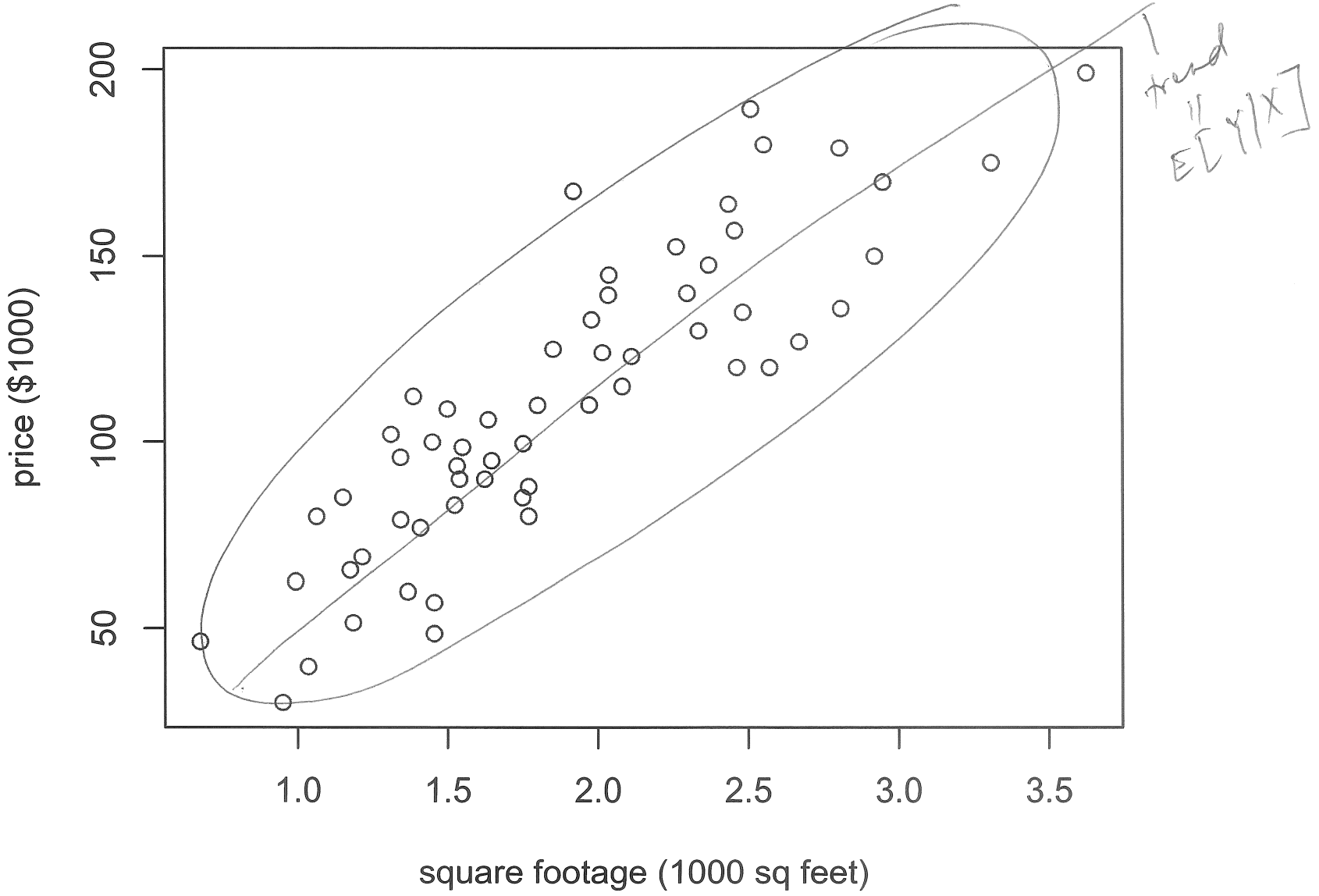
```
> head(house)
   size price
1 0.951 30.00
2 1.036 39.90
3 0.676 46.50
4 1.456 48.60
5 1.186 51.50
6 1.456 56.99
>
```

```
> str(house)
'data.frame': 58 obs. of  2 variables:
 $ size : num  0.951 1.036 0.676 1.456 1.186 ...
```

```
 $ price: num  30 39.9 46.5 48.6 51.5 ...
>

# Make a scatterplot

plot(house$size, house$price,
xlab="square footage (1000 sq feet)",
ylab="price ($1000)")
```


price ($1000)
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5
square footage (1000 sq feet)
trend
=
E[Y|X]