Department of Linguistics and Translation
香港城市大學
City University of Hong Kong

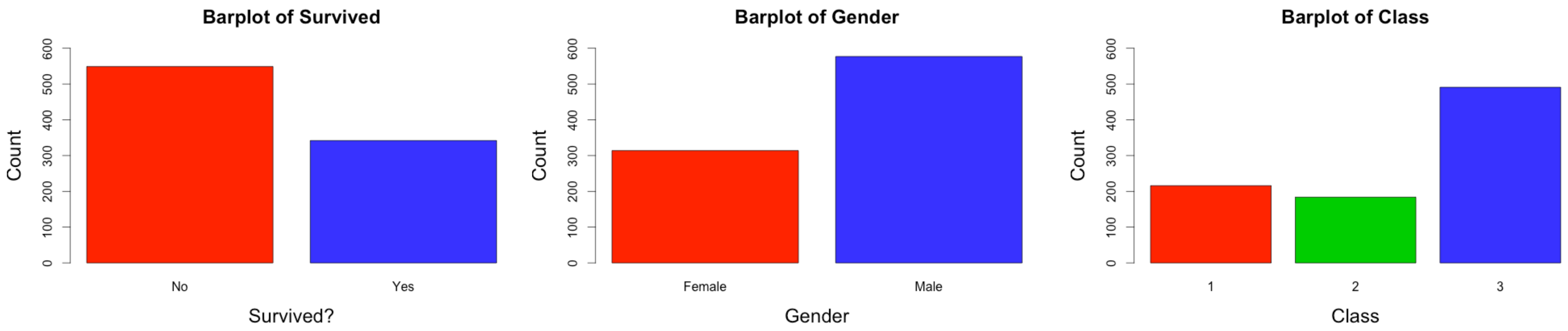
Fundamentals of Statistics for Language Sciences LT2206

Jixing Li

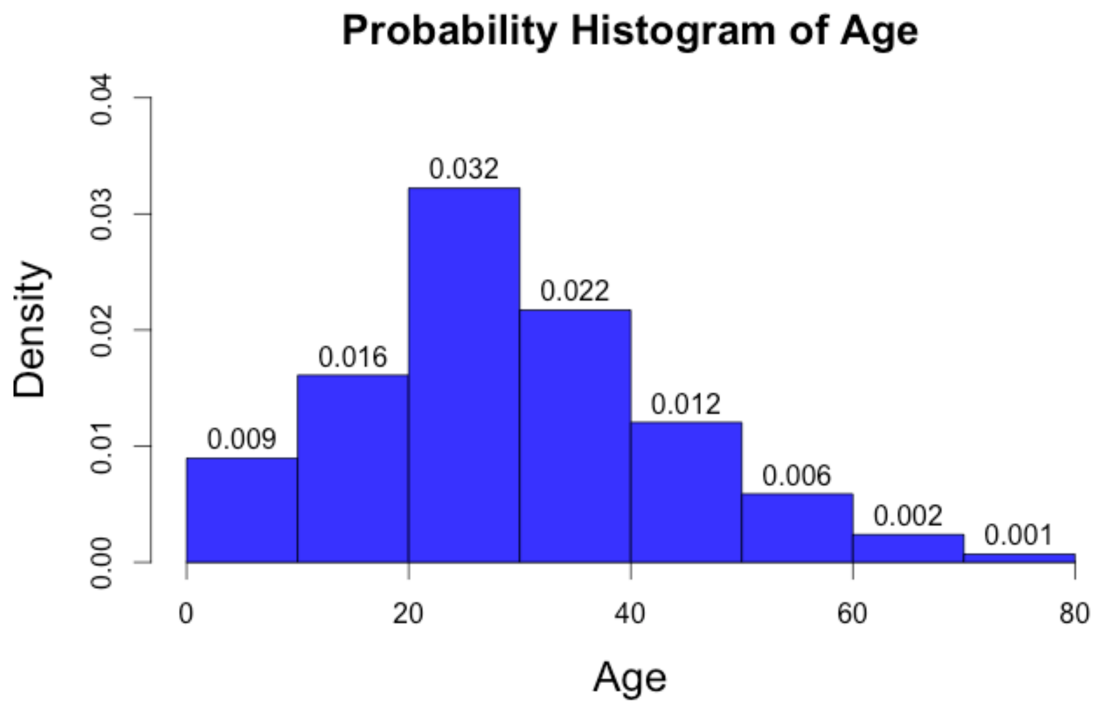
Lecture 4: Hypothesis testing

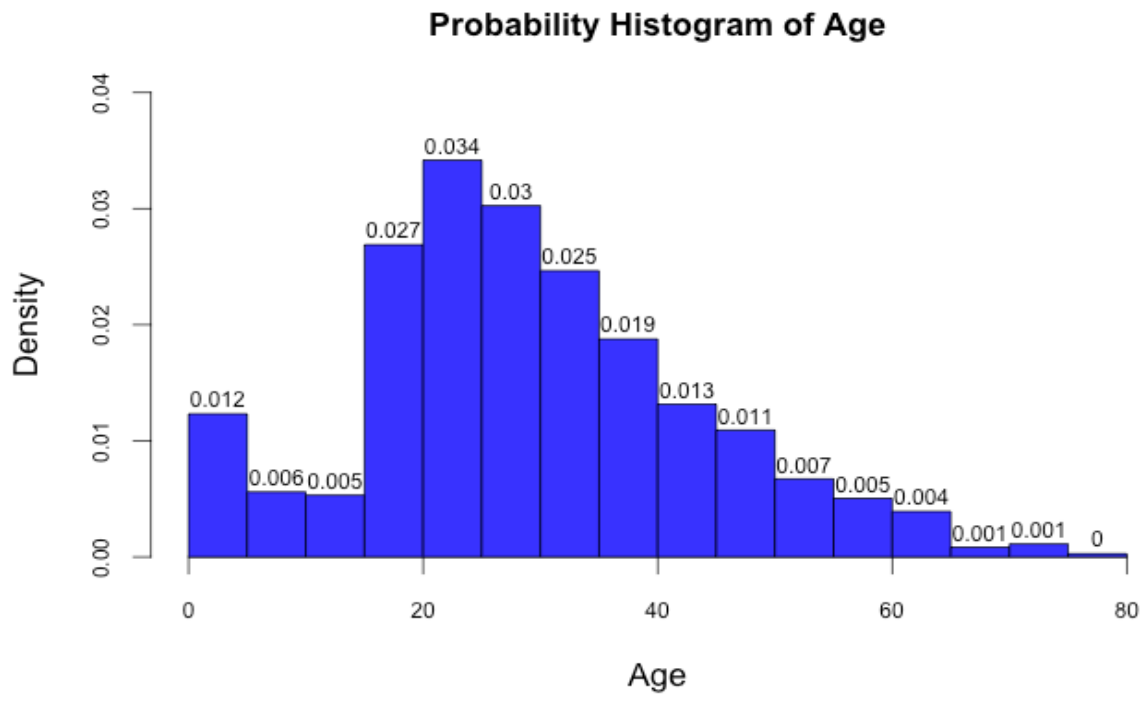
Slides adapted from Cecilia Earls

Bar plots

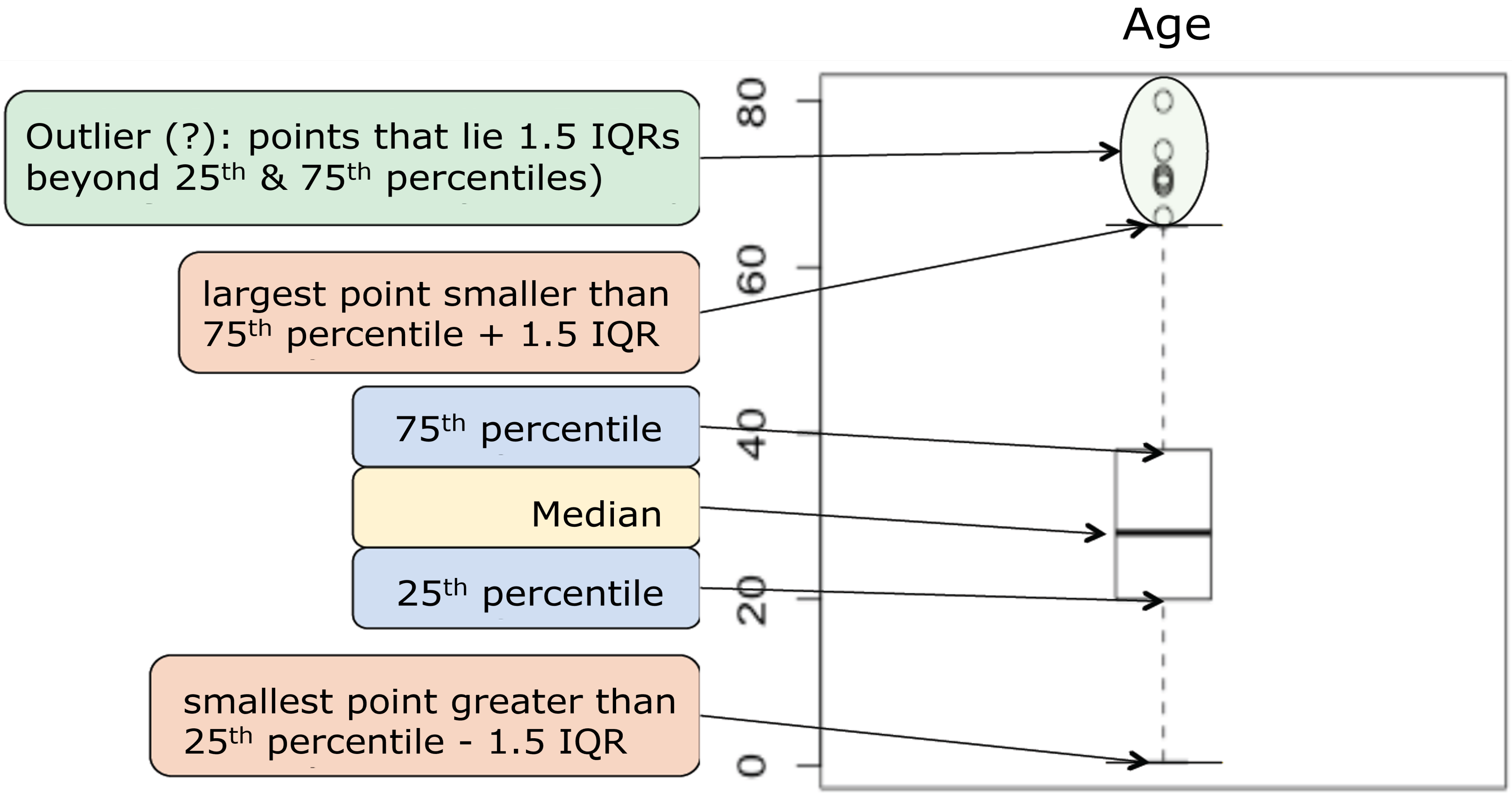

Barplot of Survived
Count
0
100
200
300
400
500
600
No
Yes
Survived?
Barplot of Gender
Count
0
100
200
300
400
500
600
Female
Male
Gender
Barplot of Class
Count
0
100
200
300
400
500
600
1
2
3
Class

Histograms

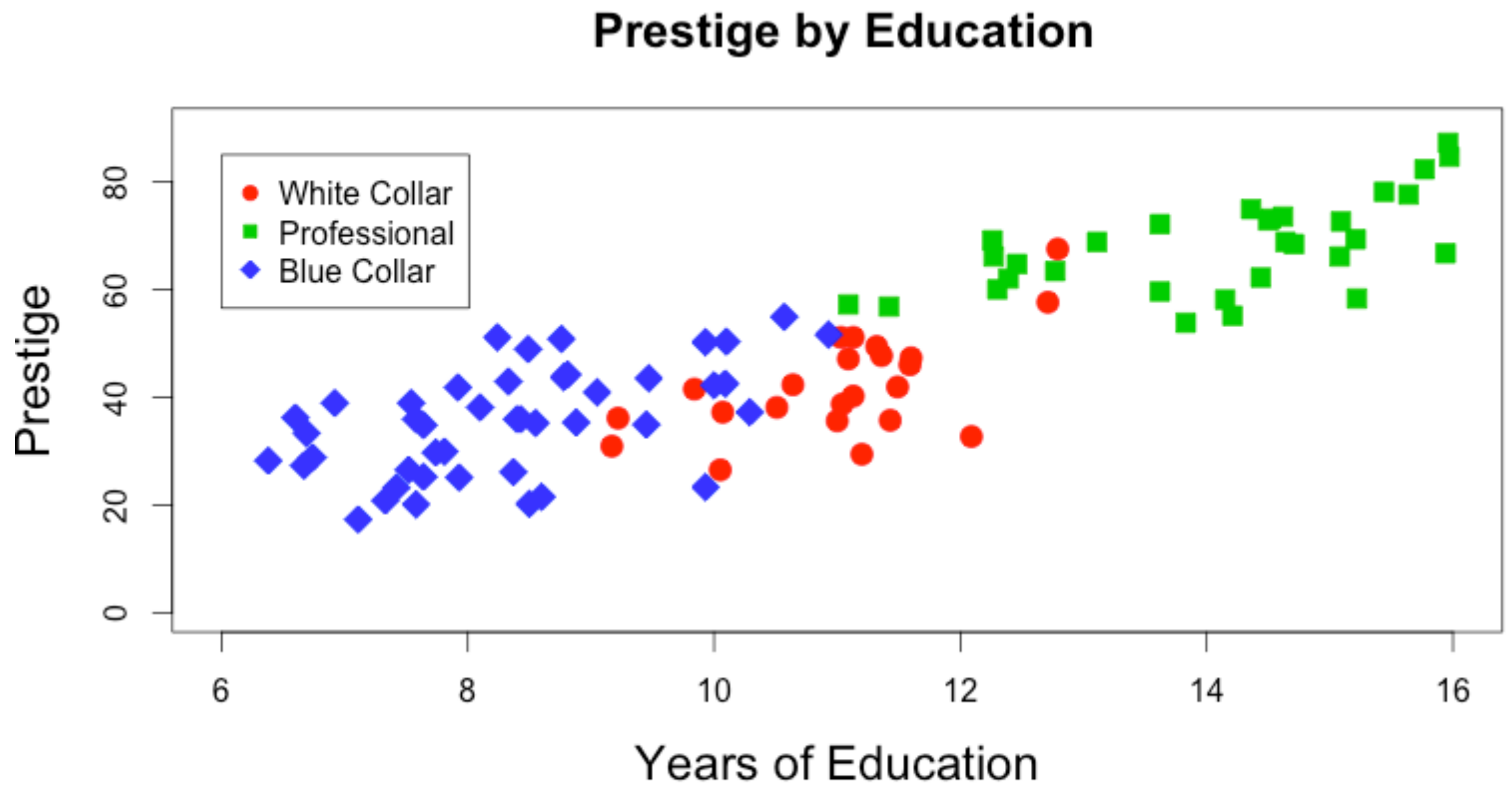

Probability Histogram of Age
Density
Age
0.009
0.016
0.032
0.022
0.012
0.006
0.002
0.001
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80


Probability Histogram of Age
Density
Age
0.012
0.006
0.005
0.027
0.034
0.03
0.025
0.019
0.013
0.011
0.007
0.005
0.004
0.001
0.001
0
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80

Boxplots


Age
Outlier (?): points that lie 1.5 IQRs beyond 25th & 75th percentiles)
largest point smaller than 75th percentile + 1.5 IQR
75th percentile
Median
25th percentile
smallest point greater than 25th percentile - 1.5 IQR
80
60
40
20
0

Scatterplots



Statistical hypothesis testing

Statistical hypotheses: Statements made about a specific value of a population parameter (e.g., μ).

Hypothesis test: Statistical method for evaluating the degree to which evidence favors (or does not favor) the alternative hypothesis over the null hypothesis.

- **H_0:** null hypothesis ("nothing going on")
- **H_a:** alternative hypothesis ("something going on")

Tests for a population mean

Let μ_0 be a specific value of the population mean of direct interest, then we can test **one** of the following three sets of hypotheses:

H_0: $\mu=\mu_0$ vs. H_a: $\mu\neq\mu_0$ Two-sided tests

H_0: $\mu\leq\mu_0$ vs. H_a: $\mu>\mu_0$

One-sided tests

H_0: $\mu\geq\mu_0$ vs. H_a: $\mu<\mu_0$

Goal: Create objective framework in which we can

- Make a decision about H_0 vs. H_a using data (e.g., $\overline{X_n}$)
- Evaluate whether the observed value of our statistic (e.g., $\overline{X_n}$) seems reasonable if H_0 is indeed true.

Idea: Use the known behavior of $\overline{X_n}$ under SRS (Simple Random Sampling)

- Construct sampling distribution of $\overline{X_n}$ under H_0 (e.g., with $\mu = \mu_0$). This gives us information on "likely" values of sample mean seen under SRS, assuming $\mu = \mu_0$.
- Use this distribution to evaluate how likely we would be to observe a value more extreme than that obtained from SRS, $\overline{X_n}$, assuming $\mu = \mu_0$.

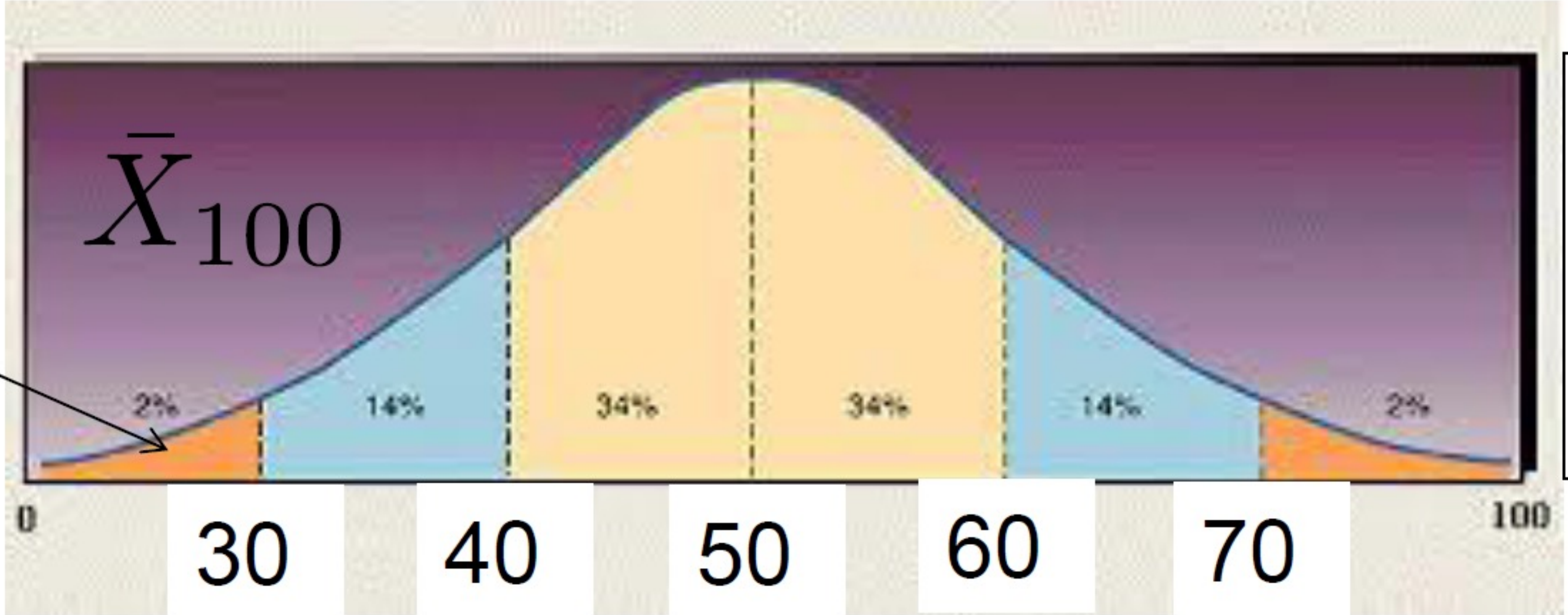
$X_i, i = 1, \dots, 100, iid, E(X_i) = \mu; SD(X_i) = 100$

We know under certain conditions…

$$\bar{X}_{100} \sim N(\mu, 100^2/100)$$

Assume! $\mu = 50$ Then: $\bar{X}_{100} \sim N(50, 100^2/100)$ under H0

$P(\bar{X}_{100} < 30)$ = 0.02



μ_0 = 50

Assumed mean under H_0

If H_0 is true, is $\bar{x}_{100} = 25$ a likely value?

How do we make a formal decision about H_0?

Decision Rules, Type I & Type II Errors, and Rejection Regions

To structure the decision making process, we need:

1. A decision rule that allows us to decide whether the observed value of $\overline{X_n}$ should be judged consistent with H_0 (i.e., with $\mu = \mu_0$) or supportive of H_a over H_0.

2. A way to include the possibility of making an error in either direction, and to assess its impact.

- Data → decision between: Reject H_0 or Do not reject H_0
- View H_0 as the null situation "nothing is going on":

↓

standard assumption is true

- View H_a as the opposite situation "something is going on"

↓

standard assumption not true – usually the research hypothesis

- Two possible types of error: Type I & Type II
  - Type I error: Reject H_0 when H_0 is actually true – "false positive"
  - Type II error: Do not reject H_0 when H_0 is actually false, that is, when H_a is true – "false negative"

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

- We want: P(Type I error) ≈ 0 and P(Type II error) ≈ 0

  How can we approach this? How well can we do?
- For a hypothesis test:
  - We specify (control) the Type I error rate, i.e., limit the "false positive" rate.
  - After fixing Type I error rate, Type II error rate can often be reduced by increasing the sample size (reduces variability)
- Type I error rate := significance level = size of test: usually denoted by $\alpha = 0.05$:

$$P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0 \text{ true}) = 0.05$$

So under H_0, this test mistakenly rejects H_0 5% of the time.

- Smaller $\alpha \rightarrow$ less chance of false positive result
- Common choices: α = 0.05, 0.01, 0.001

- Type II error rate: Usually denoted by β. Then:

$$P(\text{Type II error}) = P(\text{fail to reject } H_0 \mid H_a \text{ true})$$

$$\text{Power} = 1 - P(\text{Type II error}) = P(\text{reject } H_0 \mid H_a \text{ true})$$

Tests with low power are basically useless. However, the researcher can often increase power by increasing sample size.

Hypothesis Testing

- First Steps:
  - Set type I error rate, α
  - H_a ($\neq$ or $>$ or $<$) & H_0 ($=$ or $\leq$ or $\geq$)
  - Determine sample size to control type II error rate
  - Collect SRS
- Determine:
  - Distribution of (standardized) test statistic under H_0
  - Using this distribution make a decision based on either:
    - Rejection Region
    - p-value
- Finish: check assumptions, draw conclusions

Hypotheses we will consider for μ:

- H_a: $\mu > \mu_0$ vs. H_0: $\mu \leq \mu_0$
- H_a: $\mu < \mu_0$ vs. H_0: $\mu \geq \mu_0$
- H_a: $\mu \neq \mu_0$ vs. H_0: $\mu = \mu_0$

μ_0 = hypothesized mean used to calculate the test statistic for all three hypotheses – it is known quantity

- H_a usually reflects what the researcher would like to know. Is something different going on than what has previously been assumed?
- Can use a rejection region or a p-value to make a decision about whether something different is going on.

The Standardized Test Statistic

Assume X_i , i = 1, .., n is a SRS from some population with mean, μ, and variance σ^2 . Under H_0 we will assume $\mu = \mu_0$. Then:

- If <u>σ is known</u>, the following is the standardized test statistic under H_0.

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

- If <u>σ is unknown</u>, the following is the standardized test statistic under H_0.

$$\frac{\bar{X}_n - \mu_0}{S/\sqrt{n}}$$

Decision making using a rejection region (RR):

- RR = range of values of the standardized test statistic for which you will reject H0 if you realized standardized test statistic is in this range. For example:

$$\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}} > 1.96 \textit{ is one possile RR}$$

- RR depends on
  - Distribution of the standardized test statistic
  - Type I error rate
  - H_a (Will you reject if $\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$ is big? Small? Big in absolute value?)

Basic principle in choosing rejection region:

- If Ha: $\mu > \mu_0$ is true, then deviations of $\bar{X}_n$ from μ_0 in the positive direction (i.e., $\bar{X}_n > \mu_0$) should be considered as evidence against H_0.
- If H_a: $\mu < \mu_0$ is true, then deviations of $\bar{X}_n$ from μ_0 in the negative direction (i.e., $\bar{X}_n < \mu_0$) should be considered as evidence against H_0.
- If H_a: $\mu \neq \mu_0$ is true, then deviations of $\bar{X}_n$ in either direction away from μ_0 are evidence against H_0.

In each case: α dictates how far $\bar{X}_n$ must be from μ before we conclude that the data provide more support for H_a.

Rejection Region Approach: σ known

1. Specify H0, Ha, & α = P(Type I error)
2. Compute Test Statistic (T.S.)
- Value depends on sample collected
- Obtain its sampling distribution under assumption that H_0 is true.

3. Determine Rejection Region (R.R.)
- Specifies both the direction(s) and magnitude of deviations from the population mean that we regard as representing evidence against H_0

4. Decision / Conclusion :
- If T.S. falls in R.R., then reject H_0 in favor of H_a; otherwise, fail to reject the null hypothesis H_0.

5. Check assumptions

Example: H_0: $\mu \leq \mu_0$
H_a: $\mu > \mu_0$

T.S. $$Z^* = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

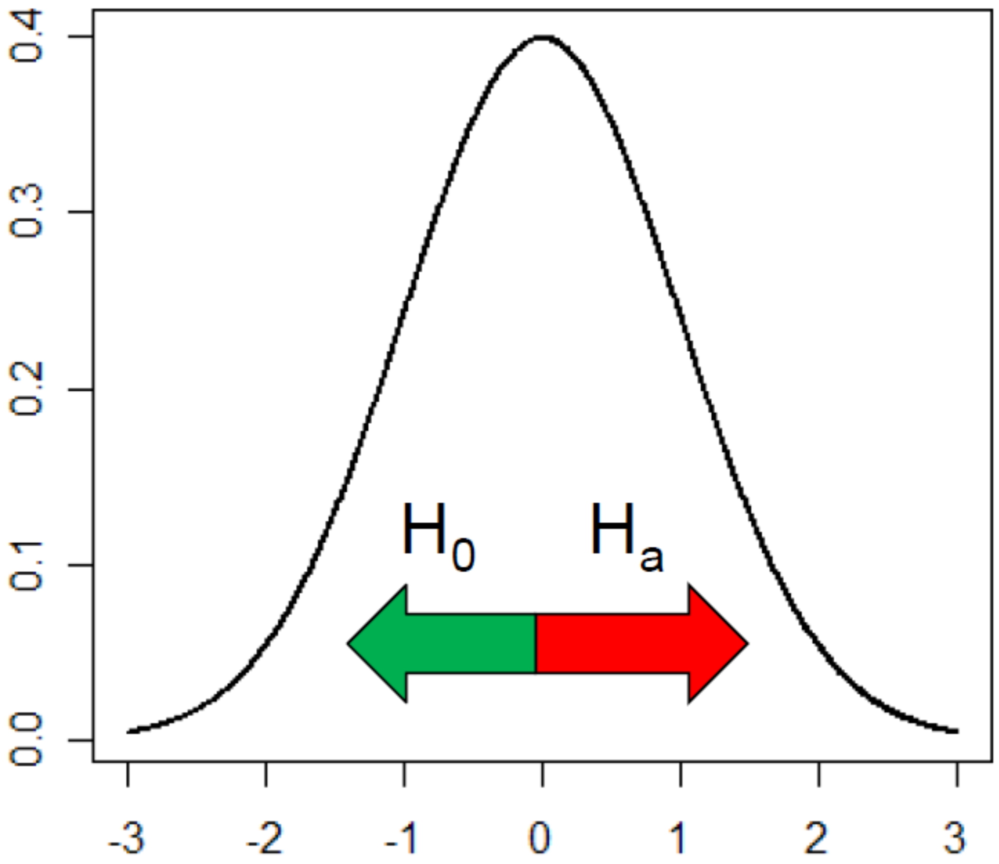
R.R. $Z^* > Z_\alpha$

For the two other forms of Ha, the relevant RRs are:
(ii) Ha: $\mu < \mu_0 \leftrightarrow Z^* < -Z_\alpha$
(iii) Ha: $\mu \neq \mu_0 \leftrightarrow |Z^*| > -Z_{\alpha/2}$

This analysis assumes the standardized test statistic has a N(0,1) distribution.

Consider the test for
H0: $\mu \leq \mu_0$ vs. Ha: $\mu > \mu_0$
In this case, values of $\bar{X}_n$ bigger than μ_0 are evidence against H0.

The realization of the standardized test statistic, $Z^* = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$, measures how many standard errors away the observed value lies from μ_0. For this H_a, anything greater than 2 is fairly good evidence against H_0. 2 is a little arbitrary; using z_α limits the type 1 error rate to be at most α.



Why does the RR given as $Z^* > Z_\alpha$ limit the type I error rate to be at most α?

Suppose α = 0.05

Suppose under H0:

$$\frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \dot{\sim} N(0,1)$$

Key idea: under repeated SRSs from a population with mean μ_0 and SD σ, we know the approximate sampling distribution of the standardized statistic. In particular, we have

$$P\left(\frac{\bar{x}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} > Z_\alpha\right) \doteq \alpha$$