Douglas Rubin

# Class 3 of Stats Bootcamp

Hypothesis Testing

# Question to think about

***You flip a coin 100 times and observe 51 heads. Is the coin fair ($p_{heads} = 0.5$) or unfair?*** Douglas Rubin

Not overwhelming evidence against being fair (in statistics terms: can't reject null hypothesis), so the coin is probably fair

***What if you had observed 90?***

Flipping 90 heads out of 100 is highly unlikely if the coin is fair, so the coin is probably unfair (the probability is in fact about $1.4 \times 10^{-17}$)

***What if you had observed 60, 65, 70?***

Maybe fair, maybe not fair. It is unclear where the threshold between fair and unfair is

**Statistical hypothesis testing** gives us a method for deciding between 2 competing hypotheses

***A subtlety to think about:***

- Have you ever seen an unfair coin in your life?
- How would this change the probability that the coin is unfair if, say, 80 heads were observed?
- Probabilistically including prior beliefs is the subject of Bayesian inference.

# Class Outline

1. Introduction to hypothesis testing by way of example: 2 sided z-test
    - null and alternative hypotheses
    - p-values and $\alpha$
    - the null distribution
    - rejecting the null hypothesis
2. General hypothesis testing procedure
3. Example 2: 1 sided t-test
4. Example 3: a non-parametric test
5. Example 4: 1 sided z-test comparing 2 means
6. Example 5: 2 sided test with a Binomial

# Example 1: Are New Yorkers fat or skinny?

- Basic 2 sided z-test

It is known that the national average weight of a 50 year-old male has a distribution with $\mu_o = 194.4$lbs with $\sigma_o = 49.3$lbs. We suspect that since New Yorkers lead such a chaotic lifestyle, that their weights may be different than the national average.

***Given the sample of 50 year-old NYC male weights from the previous class*** ($N = 1000$, $\bar{x} = 198.1 lbs, s = 54.4 lbs$), ***are the weights of New Yorkers different than they are nationally?***

# Null and alternative hypothesis

- Let's start by forumlating 2 mutually exclusive (and typically) collectively exhaustive hypotheses, so that if we reject 1, then by default we accept the other.

**Null hypothesis** ($H_o$): This is usually an uninteresting hypothesis where we typically posit that there is no effect/difference (e.g. a certain medication doesn't do better than a placebo)

**Alternative hypothesis** ($H_A$): This is the more interesting hypothesis where we posit that there $is$ an effect/difference (e.g. a certain medication $does$ do better than a placebo)

# Null and alternative hypothesis (cont.)

- In hypothesis testing we start off by assuming that the uninteresting, null hypothesis is true, and only reject it in favor of the interesting, alternative hypothesis when there is sufficient evidence to do so.
    - reaching a point where we can reject/fail-to-reject the null hypothesis is whole goal in hypothesis testing
    - if we can reject $H_o$ in favor of $H_A$, we will have shown something interesting (e.g. that the medication probably works)
- By, assuming the boring/null hypothesis $a\ priori$, and only rejecting it for the interesting/alternative hypothesis when there is sufficient evidence to do so, this is a conservative approach to accepting the interesting hypothesis

# Null and alternative hypothesis (cont.)

In our example:

$H_o$: The average weight of 50 year-old NYC males follows the national distribution

$\implies \mu_{NYC} = \mu_o, \sigma_{NYC} = \sigma_o$

$H_A$: The average weight of 50 year-old NYC males is different than the national distribution

$\implies \mu_{NYC} \neq \mu_o$

# p-values

Douglas Rubin

- What sort of "evidence" do we need in order to try to reject $H_o$ if there is sufficient "evidence" to do so?

A **p-value** (or just $p$), $very$ loosely, is the probablility, assuming the null hypothesis is true, of obtaining what was actually observed (usually a sample mean, $\bar{x}$).

- If we can manage to calculate $p$, then we are finished with the problem because:
  - if $p$ is low, we reject $H_o$ (in favor of $H_A$)
  - if $p$ is high, we cannot reject $H_o$

**Remark:**

- More precisely, a $p$-value has 2 equivalent definitions:
  - Assuming $H_o$, $p$ is the probability of observing data at least as extreme as what was actually observed.
  - $p$ is the probability of making a type I error if we set our reject/fail-to-reject threshold at the observed value, $\bar{x}$.
    - **type I error**: incorrectly rejecting $H_o$, i.e., rejecting $H_o$ when it is in fact true

# Significance levels, $\alpha$

- How low is low enough?
  - We usually set a threshold value, $\alpha$, where if $p < \alpha$ we reject $H_o$. We call this a **significance level**.
  - $\alpha = 0.05$ is a typical choice.
  - Values $< 0.05$ are sometimes used to be extra conservative.
  - It is customary to state $\alpha$ with the results of a hypothesis test, since rejection/not rejection depends on the value of $\alpha$.

For this example, let's set a threshold of $\alpha = 0.05$

# Formulating a null distribution to calculate p

Douglas Rubin

- The goal now is to try to calculate $p$ so that we can reject or fail-to-reject $H_o$
- Again, the $p$-value is the probability, assuming $H_o$ is true, of observing data at least as extreme as what was actually observed, $\bar{x}$.
- Recall, we can compute probabilities from the area under a probability distribution, so let's try to construct a probability distribution for $\bar{X}$ assuming $H_o$ is true.

**Recall:**

- The Gaussian probability density is given by: $P(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$.
- If the probability density function for a random variable, Y, is a Gaussian with mean $\mu$ and variance $\sigma^2$, we denote this by $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

# Formulating a null distribution to calculate p (cont.)

Douglas Rubin

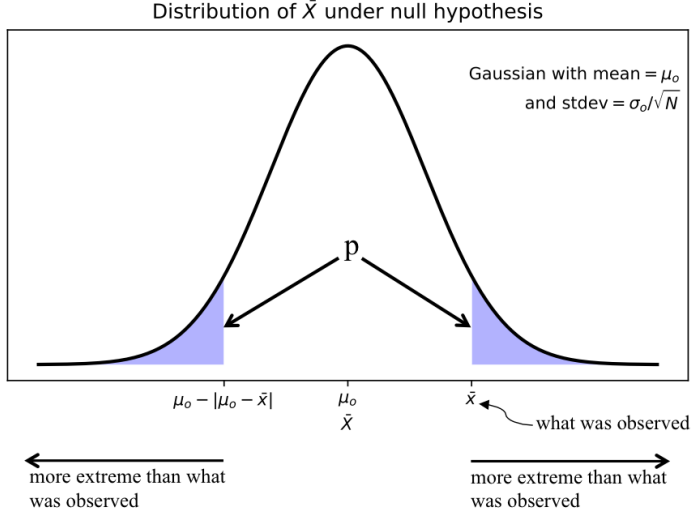
In our example, assuming the null hypothesis is true:

$\mu_{NYC} = \mu_o$ and $\sigma_{NYC} = \sigma_o$

We know from the Central Limit Thereom that:

$$\bar{X} \sim \mathcal{N}\left(\mu_{NYC}, \frac{\sigma^2_{NYC}}{N}\right) = \mathcal{N}\left(\mu_o, \frac{\sigma^2_o}{N}\right)$$

Thus, if $H_o$ is true, the probability, of observing data at least as extreme as what was actually observed, $\bar{x}$ is given by:



# Calculating $\mathbf{p}$ and accepting/rejecting $\mathbf{H_o}$

Douglas Rubin

- If the null distribution is Gaussian, the shaded area under the curve, $p$, is simply given by integrating a Gaussian over the approriate range.
- This can be found by calculating $z = \frac{\bar{x}-\mu_o}{\sigma_o/\sqrt{N}}$ (called a **z-score**) and looking up the corresponding $p$-value in a Z table, or by plugging z into statistical software (proof on next slide).

Completing our example:

$$z = \frac{\bar{x}-\mu_o}{\sigma_o/\sqrt{N}} = \frac{198.1lbs-194.4lbs}{49.3lbs/\sqrt{1000}} \cong 2.37$$

2-sided $z$-test: $p(2.37) \cong 0.018$

$$\Longrightarrow$$

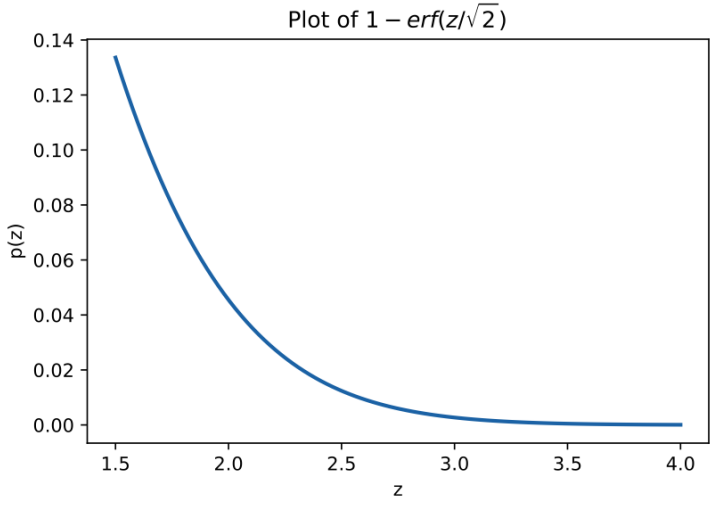
$p$ is less than our pre-selected $\alpha$ level of 0.05. We therefore **reject** the null hypothesis that $\mu_{NYC} = \mu_o$ (at a signficance level of $0.05$) in favor of the alternative hypothesis that the weights of 50 year-old NY males are different than the national average.

# 2-sided p-values for a Gaussian null distribution (optional)

Douglas Rubin

$$
\begin{aligned}
p &= \frac{2}{\sqrt{2\pi}\left(\sigma_o/\sqrt{N}\right)} \int_{\bar{x}}^{\infty} e^{-\frac{(\bar{X}-\mu_o)^2}{2(\sigma^2/N)}} d\bar{X} \\
&= \frac{2}{\sqrt{2\pi}} \int_{z}^{\infty} e^{-\frac{Z^2}{2}} dZ \\
&= 1 - \mathrm{erf}(z/\sqrt{2})
\end{aligned}
$$

- I have changed integration variables with $Z = \frac{\bar{X}-\mu_o}{\sigma_o/\sqrt{N}}$. The erf function is an integral over a certain domain of a Gaussian and must be computed numerically.



# Summary of how to conduct a hypothesis test

Douglas Rubin

1. Pick a statistic to test
2. Formulate $H_o$ and $H_A$
3. Assuming null hypothesis true:
    A. Identify distribution under null hypothesis
    B. Chose an alpha threshold level
    C. Calculate p and compare to alpha to reject or fail to reject

        - determine whether to use a 1-sided or 2-sided test (more to be said about this)
        - if the null distribution is Normal, we can compute $p$ with a z-test
        - if the null distribution is a t-distribution, we can compute $p$ with a t-test
        - other parametric null distributions are possible, and even we cannot identify the proper distribution, we can always resort to non-parametric tests

In Example 1:

1. $\bar{X}$ (sample mean weight)
2. $H_o : \mu_{NYC} = \mu_o, \sigma_{NYC} = \sigma_o$ and $H_A : \mu_{NYC} > \mu_o$
3. assuming $H_o$:
    A. $\bar{X} \sim \mathcal{N}\left(\mu_o, \frac{\sigma_o^2}{N}\right)$
    B. $\alpha = 0.05$
    C. 2-sided z-test, $p = 0.045 < \alpha \implies$ reject $H_o$

# A note on 1-sided vs. 2 sided tests

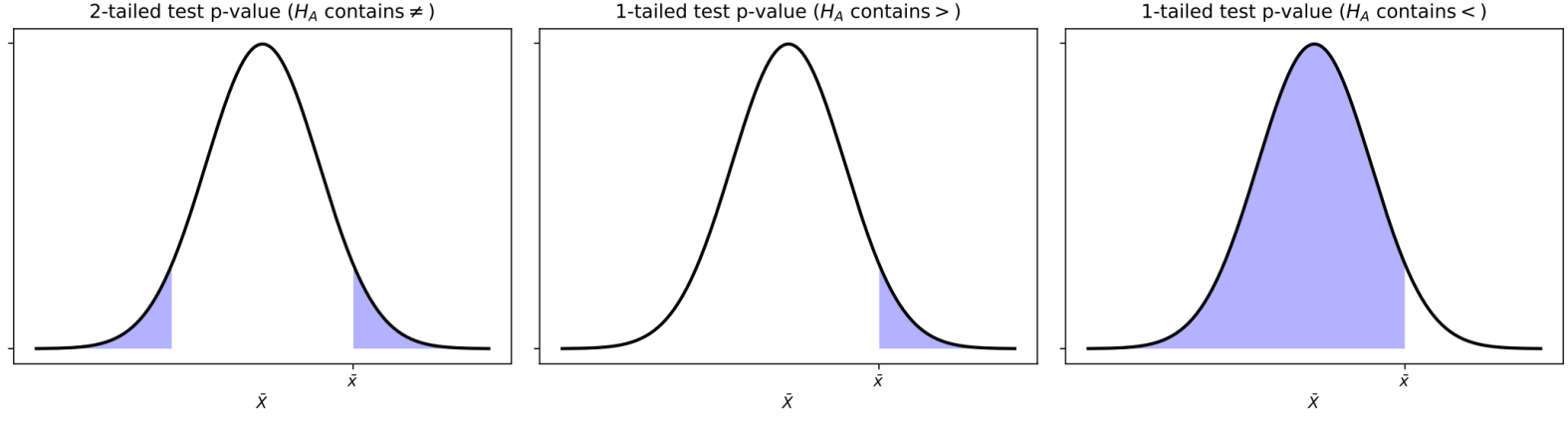
- For $H_A : \mu \neq \mu_o$ ($\mu \neq \mu_o$ is the same as $\mu < \mu_o$ or $\mu > \mu_o$) we use a 2-sided test

Douglas Rubin

  - Very high **or** very low values of $\bar{x}$ is evidence in favor of $H_A$ over $H_o$

- For $H_A : \mu > \mu_o$ we use a 1-sided test
  - Very high values of $\bar{x}$ is evidence in favor of $H_A$ over $H_o$

- For $H_A : \mu < \mu_o$ we use a 1-sided test
  - Very low values of $\bar{x}$ is evidence in favor of $H_A$ over $H_o$




- Another, potentially more satisfying explanation as to when it is appropriate to use a 1-sided vs. 2-sided test can be given when viewing a $p$-value in terms of its type 1 error definition

# Example 2: Are CEOs tall?

- 1 sided t-test comparing means

The average male height in America follows a normal distribution with $\mu_o = 170$cm. We suspect that being tall may be advantageous for whether one attains a c-level position in business. We therefore randomly sample 9 c-level men and measure their heights, obtaining $\bar{x} = 174.8$cm and $s = 8.25$cm.

***Are c-level men taller on average?***

# Example 2

Douglas Rubin

**test statistic:** $\bar{X}$ (sample mean height)

**hypotheses:**

$H_o: \mu_{CL} = \mu_o$

$H_A: \mu_{CL} > \mu_o$

## Assuming $\mathbf{H_o}$

**null Distribution:**

- Because of the CLT, $\bar{X} \sim \mathcal{N}\left(\mu_o, \frac{\sigma_o^2}{N}\right)$
- Since we do not know $\sigma_o$, we will have to approximate it with $s$.
- As in the previous lecture, since the sample size $N$ is small the appropriate distribution to use is a t-distribution with $df = N - 1 = 8$.

# Example 2 (cont.)

**significance level:** $\alpha = 0.05$

**p-value:**

$H_A$ has $>$ $\implies$ 1-sided test

To calculate $p$, as with the Gaussian case, we can "standardize" $\bar{x}$ and look up the corresponding $p$-value in a t-table.

$$t = \frac{\bar{x}-\mu_o}{\sigma_o/\sqrt{N}} \approx \frac{\bar{x}-\mu_o}{s/\sqrt{N}} = \frac{174.8cm-170cm}{8.25cm/\sqrt{9}} = 1.75$$

1-sided t-test: $p(1.75) \cong 0.12$

$\implies$ We cannot reject the null hypothesis at a level of $\alpha = 0.05$. However, since the $p$-value is still quite small, this implies that there is evidence that c-level males are taller than the average. Acquiring more data points (and hence getting a better estimate of the population mean) could lower the $p$-value below $\alpha$.

# Example 3: Is the A train slower than the C?

- A non-parametric hypothesis test
  - useful when you don't know the null distribution, or suspect it may not be normal or t-distributed

Every day Sergey takes either the A train or C train from Brooklyn to Times Square. Even though they are approximately the same distance and number of stops, Sergey believes that the A train is less efficient and hence has a longer median travel time. Being a slightly nerdy data scientist, Sergey decides to collect data to test his hypothesis. He randomly chooses 15 days to ride the A train and collects a dataset with a sample median of $\widehat{x}_A = 28.7$mins. He also randomly chooses 20 days to ride the C train and collects a dataset with a sample median of $\widehat{x}_C = 22.3$mins.

***Is the median travel time of the A train longer than that of the C?***

# Example 3

**test statistic:** $\widehat{X}_A - \widehat{X}_C$

**hypotheses:**

$H_o$: $\text{med}_\text{A} - \text{med}_\text{C} = 0$, and further the travel times for A and C follow the same distribution.

$H_A$: $\text{med}_\text{A} > \text{med}_\text{C}$

- $\text{med}_\text{A}$ and $\text{med}_\text{C}$ denote the true, population medians.

# Example 3 (cont.)

Douglas Rubin

## Assuming $\mathbf{H_o}$

**Null Distribution:**

- need the null distribution for the difference in sample medians, $\widehat{X}_A - \widehat{X}_C$
- CLT only applies to sample means, **not** sample medians (although, one can construct something like a CLT for medians), so it is unclear what the null distribution should be.

- we can use a non-parametric, **bootstrap** approach:
  - many new "fake" sample datasets are generated from the sample dataset (using sampling with replacement)
  - $\widehat{X}_A - \widehat{X}_C$ is computed for each new "fake" sample
  - the distribution of $\widehat{X}_A - \widehat{X}_C$ is given by the normalized frequency distribution of $\widehat{X}_A - \widehat{X}_C$ generated by the "fake" samples

- Assuming $H_o$, the times for A and C follow the same distribution and we can merge the A and C datasets into one large dataset.

# Example 3, Bootstrapping (cont.)

Douglas Rubin

$N_A = 2$

$N_C = 3$

$\text{Data}_A = \{3, 8\}$

$\text{Data}_C = \{6, 5, 5\}$

Data = {3, 8, 6, 5, 5}

"Data"$_A^{(1)}$ = {8, 8}
"Data"$_C^{(1)}$ = {6, 3, 5}

$\hat{x}_A^{(1)} - \hat{x}_C^{(1)} = 3$

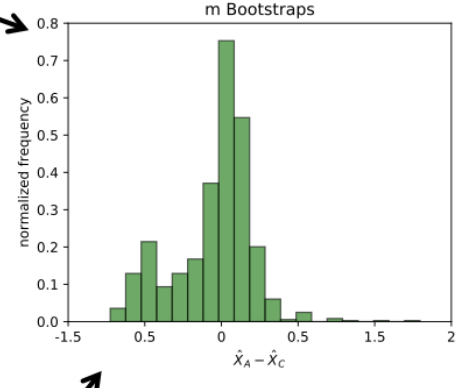
$(8 - 5 = 3)$

"Data"$_A^{(2)}$ = {5, 5}
"Data"$_C^{(2)}$ = {6, 8, 3}

$\hat{x}_A^{(2)} - \hat{x}_C^{(2)} = -1$

"Data"$_A^{(m)}$ = {6, 5}
"Data"$_C^{(m)}$ = {5, 6, 5}

$\hat{x}_A^{(m)} - \hat{x}_C^{(m)} = 0.5$



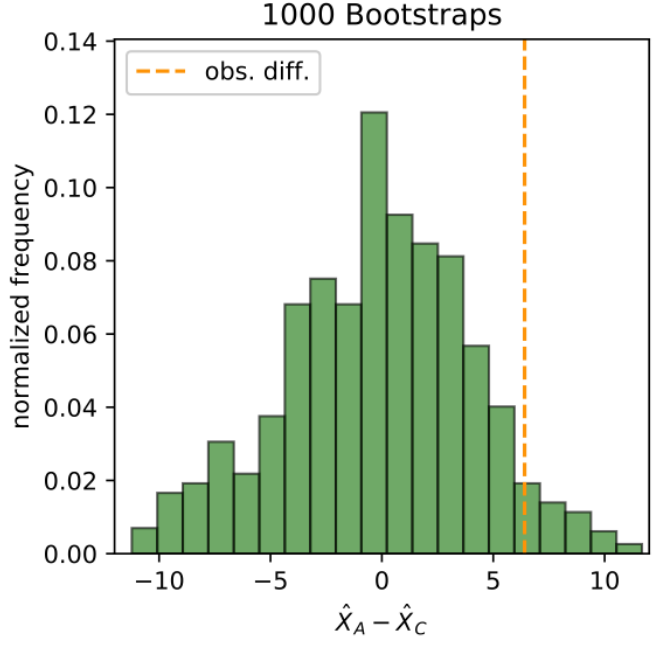
# Example 3, Bootstrapping (cont.)

- Roughly speaking, the intuition behind why bootstrapping should work is:

Douglas Rubin

Since the sample data was sampled from the population, the sample distribution roughly has the same properties as the population distribution. Therefore if we randomly sample from from the sample data, we may create new "fake" sample datasets that would roughly approximate new, randomly selected samples from the actual population.

- The null distribution after bootstrapping the $A \cup C$ data:



# Example 3 (cont.)

**significance level:** $\alpha = 0.05$

**p-value:**

$H_A$ has $>$ $\implies$ 1-sided test

The $p$-value is just the fraction of area in the distribution to the right of the orange dashed line (the observed difference in medians).

$p \cong 0.049$

$\implies$

Sergey can reject the null hypothesis at a level of $\alpha = 0.05$ that the median travel times for the A and C train are equal if favor of the hypothesis that the median travel time for A is longer.

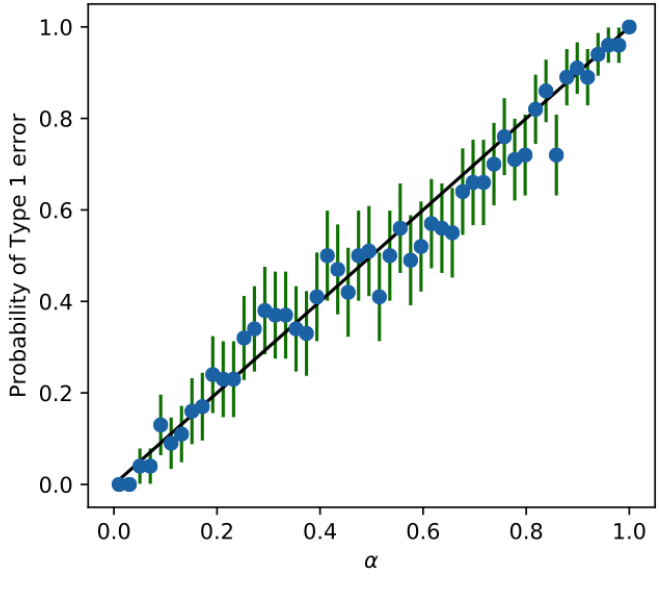
# Testing this bootstrapping procedure (advanced)

To test the mathematical machinery behind hypothesis testing, as well as the bootstrapping procedure I followed above for calculating the null distribution, I do the following:

Douglas Rubin

I sample both the A and C train samples from the same population distribution (i.e., the null hypothesis is true) with a population median value of 25.68 (I sample from a Gamma distribution). According to hypothesis testing theory, $\Pr(\text{Type I error}) \leq \alpha$ (where a Type I error is defined as rejecting $H_o$ when in fact it was true). I therefore fix $\alpha$ and randomly sample a sample of $N = 15$ and another sample of $N = 20$. I then run this through my bootstrapping procedure to calculate a $p$-value to decide whether to reject or accept the null hypothesis. For each value of $\alpha$, I repeat this 100 times to estimate the probability of making a type I error (which is the fraction of the 100 iterations that the null hypothesis was rejected). I perform this for a range of $\alpha$ values and plot the results below. The error bars are the standard errors at a level of 95%.

I also plot the $45^\circ$ line. Accoring to hypothesis testing theory, $\Pr(\text{Type I error})$ should be less than or equal to $\alpha$, so the blue points should be at, or below the solid line. We see that this procedure really doesn't do such a bad job. The data points follow the line pretty closely, and for the most part are below it.



# Example 4: Is NYC fatter than LA?

- 1 sided z-test comparing means

Data similar to the NYC data ($N_{NY} = 1000, \bar{x}_{NY} = 198.1$lbs, $s_{NY} = 54.4$lbs) is collected for 50 year-old males in LA ($N_{LA} = 2000, \bar{x}_{LA} = 196.0$lbs, $s_{LA} = 50.3$lbs).

***Are New Yorkers fatter than Los Angelians on average?***

# Example 4

**test statistic:** $\bar{X}_{NY} - \bar{X}_{LA}$

Douglas Rubin

**hypotheses:**

$H_o$: on average, NYC and LA are the same weight ($\mu_{NY} = \mu_{LA}$)

$H_A$: on average, NYC is fatter than LA ($\mu_{NY} > \mu_{LA}$)

## Assuming $\mathbf{H_o}$

**null Distribution:**

We know the distributions of both $\bar{X}_{NY}$ and $\bar{X}_{LA}$ immediately from the CLT:

$$\bar{X}_{NY} \sim \mathcal{N}\left(\mu_{NY}, \frac{\sigma^2_{NY}}{N_{NY}}\right)$$

$$\bar{X}_{LA} \sim \mathcal{N}\left(\mu_{NY}, \frac{\sigma^2_{LA}}{N_{NY}}\right)$$

# Example 4 (cont.)

**Fact**: for 2 (independant) Gaussian distributed random variables the distribution of the difference is a Gaussian where the means subtract and the variances add. Therefore:

$$\bar{X}_{NY} - \bar{X}_{LA} \sim \mathcal{N}\left(\mu_{NY} - \mu_{LA}, \frac{\sigma_{NY}^2}{N_{NY}} + \frac{\sigma_{LA}^2}{N_{LA}}\right) \approx \mathcal{N}\left(\mu_{NY} - \mu_{LA}, \frac{s_{NY}^2}{N_{NY}} + \frac{s_{LA}^2}{N_{LA}}\right)$$

Under $H_o, \mu_{NY} - \mu_{LA} = 0$, so the null distribution is:

$$\bar{X}_{NY} - \bar{X}_{LA} \sim \mathcal{N}\left(0, \frac{s_{NY}^2}{N_{NY}} + \frac{s_{LA}^2}{N_{LA}}\right)$$

# Example 4 (cont.)

**significance level:** $\alpha = 0.05$

Douglas Rubin

**p-value:**

$H_A$ has $>$ $\implies$ 1-sided test

The null distribution is Normal, so this is just a 1-sided $z$-test, so we must calculate the $z$ statistic, $z = (\bar{x} - \mu_o)/(\sigma_o/\sqrt{N})$, where, after a little thought you can determine that:

$$\bar{x} \to \bar{x}_{NY} - \bar{x}_{LA} = 198.1lbs - 196.0lbs = 2.1lbs$$

$$\mu_o \to 0$$

$$\sigma_o/\sqrt{N} \to \sqrt{\frac{s_{NY}^2}{N_{NY}} + \frac{s_{LA}^2}{N_{LA}}} = \sqrt{\frac{54.4lbs^2}{1000} + \frac{50.3^2}{2000}} = 2.06lbs$$

$$z = \frac{\bar{x}-\mu_o}{\sigma_o/\sqrt{N}} = \frac{2.1lbs-0lbs}{2.06lbs} \cong 1.02$$

1-sided $z$-test: $p(1.02) \cong 0.15$

$\implies$ We cannot reject the null hypothesis at a level of $\alpha = 0.05$. However, since the $p$-value is still quite small, this implies that there is evidence that NY is fatter than LA. Acquiring more data points (and hence getting a better estimate of the population means) could lower the $p$-value below $\alpha$.

# Example 5: Is a coin fair?

- 2-tailed test with a Binomial null distribution

***How do you determine whether a coin is fair or not by flipping it 100 times?***

# Binomial Distribution

- Let's say we flip a fair coin $n$ times and we want to know the probability of observing $k$ heads. For example, we flip a coin twice ($n = 2$), and we would like to know the probability of having observed just 1 head ($k = 1$) out of the 2 flips.
- We can enumerate the probabilities for small $n$:



$\mathbf{n = 1}$: $\{H, T\}$

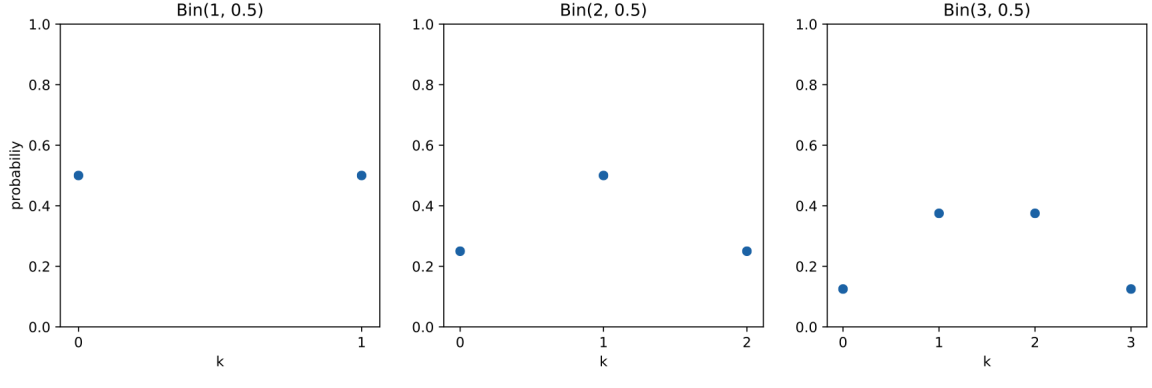
$$\implies \Pr(k = 0) = \frac{1}{2},\ \Pr(k = 1) = \frac{1}{2}$$

$\mathbf{n = 2}$: $\{HH, HT, TH, TT\}$

$$\implies \Pr(k = 0) = \frac{1}{4},\ \Pr(k = 1) = \frac{2}{4} = \frac{1}{2},\ \Pr(k = 2) = \frac{1}{4}$$

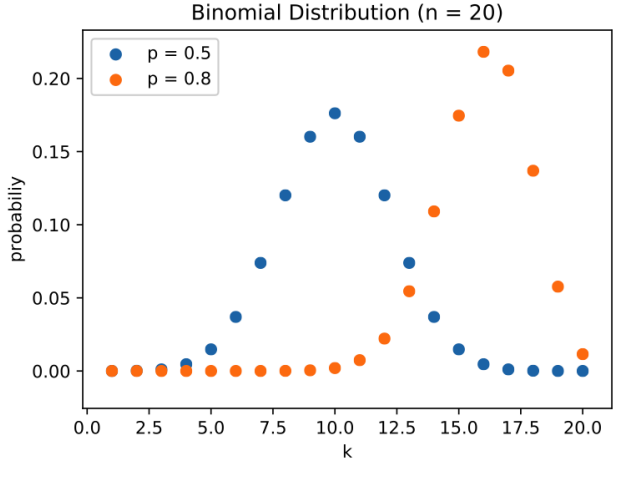
$\mathbf{n = 3}$: $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

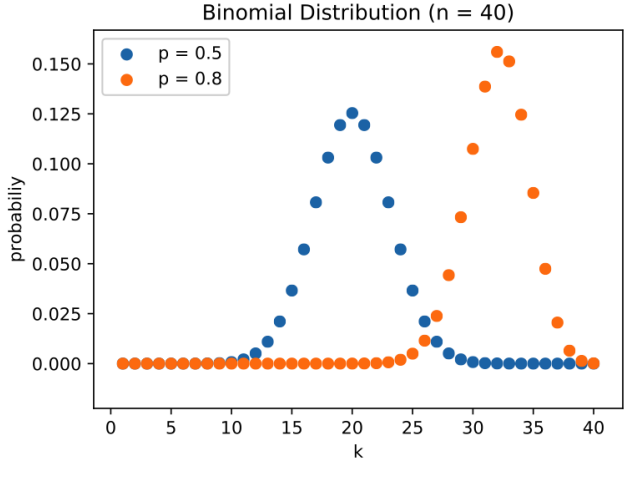
$$\implies \Pr(k = 0) = \frac{1}{8},\ \Pr(k = 1) = \frac{3}{8},\ \Pr(k = 2) = \frac{3}{8},\ \Pr(k = 2) = \frac{1}{8}$$



# Binomial Distribution

- In general the probabilities for observing $k$ heads out of $n$ flips of a coin with probability $\phi$ (not necessarily fair) is given by the Binomial distribution.
- If a random variable is distributed as a Binomial, we denote this as $k \sim \text{Bin}(n, \phi)$.



# Example 5

**test statistic:** $n_H$

**hypotheses:**

$H_o$ : coin is fair $\implies$ $n_H \sim \text{Bin}(100, 0.5)$

$H_A$ : coin is not fair $\implies$ $n_H \sim \text{Bin}(100, p_{heads})$, where $p_{heads} \neq 0.5$

## Assuming $\mathbf{H_o}$

**null Distribution:**

By assumption, the null distribution is $n_H \sim \text{Bin}(100, 0.5)$

**significance level:** $\alpha = 0.05$

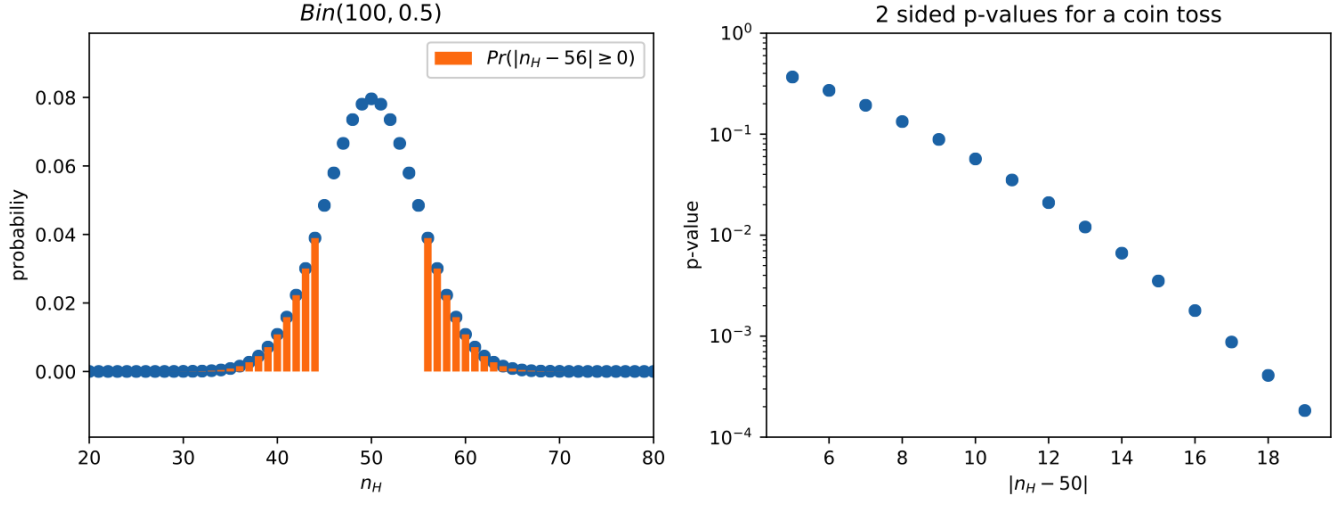
# Example 5 (cont.)

**p-value:**

Douglas Rubin

$H_A$ has $\neq \implies$ 2-sided test

Let's say we flip a coin 100 times and observe 56 heads. The 2 sided $p$-value for the binomial is given by summing the orange bars in the following figure.




$\implies p = 0.27$, so we cannot reject the null hypothesis that the coin is fair at a level of $\alpha = 0.05$.

- Note, as shown in the figure on the right, the 2-sided $p$-value for the binomial becomes very small for $|n_H - 50| \gtrsim 15$. It is therefore highly unlikely to see values of $n_H \gtrsim 65$ or $\lesssim 35$ if the coin is fair.