Statements about Means

M Loecher

Comparing distributions

"Comparing distributions for equality of means" is a process we (often unknowingly) perform all the time, e.g.:

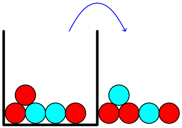
1. September 2016 seemed much warmer than "average".
2. Women tend to earn less than men.
3. Obtaining a master degree typically yields a higher salary.
4. Thomas Mueller is better at penalty shots than Lionel Messi.
5. Traffic on Mondays is the worst.
6. Colorful, attention grabbing banners on Web pages lead to more clicks.
7. The ability to concentrate is lower after drinking 4 cups of coffee.

In all of these situations, we have two groups of data - let us call them x_A and x_B - which fluctuate around their respective **true** means μ_A and μ_B, often substantially.

Why is it then that we cannot simply compare the "outcomes" and conclude that $\mu_A > \mu_B$?

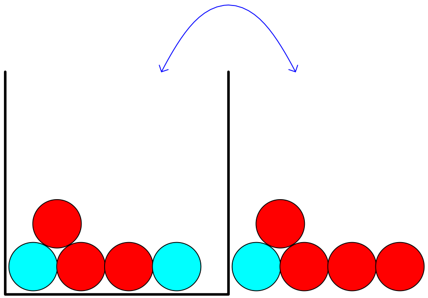
No stats needed:

1. dice: "die 2 yielded a higher number than die 1"
2. Avg. Temperature in Berlin in 2015 versus 1950
3. urn: drawing **without replacement** (until depleted)
   - "There are more red than blue marbles in the urn"
   - "The proportion of red marbles is exactly 3/5"



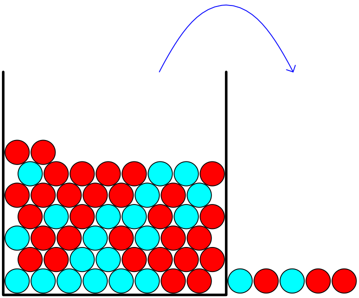
With replacement

- "There are more red than blue marbles in the urn"
- "The proportion of red marbles is exactly 4/5"



Without replacement

- "There are more red than blue marbles in the urn"
- "The proportion of red marbles is exactly 3/5"



What is a "sample" mean?

We typically **never observe the mean** μ (hence often called a "hidden" or "latent" variable), instead we observe data from distributions with **population means** μ_i.

Random Walk Simulation

Meet your good friends: **sample**, **rnorm**, **runif**, **set.seed**

1. Toss 10^5 coins $c(-1, 1)$ and store in a $1000x100$ matrix
2. Compute the cumulative sum for each column and plot
3. Compute the cumulative mean for each column and plot
4. What are the mean and variance of the 1st and last rows (both in theory and empirically)?

(New commands needed: **matrix**, **apply**, **cumsum**, **for**)

Adding/Subtracting random variables

- Compute the variance for each column from the two cumulative measures
- Udacity: golfing
- Stocks: Markowitz portfolio theory

$$x_\Sigma = \sum_{i=1}^{N} x_i \Rightarrow \sigma_\Sigma^2 = \sum_{i=1}^{N} \sigma_i^2 = N \cdot \sigma^2$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \Rightarrow \sigma_{\bar{x}}^2 = \sum_{i=1}^{N} \frac{\sigma_i^2}{N^2} = \frac{\sigma^2}{N}$$

$$\Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Inference

All we have are the sample means e.g. $\bar{x}_A > \bar{x}_B$! We nevertheless want to make statements and draw conclusions such as $\mu_A > \mu_B$. That daring step is called **statistical inference**.

$$\bar{x} \Rightarrow \mu$$

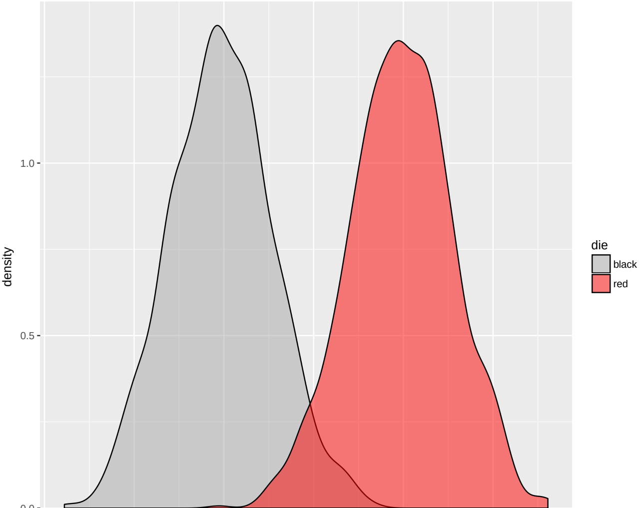
Let us play a game with two dice, one regular die and one "biased" die where we replaced the 1 with a 7. Clearly the average for the latter is greater than the former:
$(2+3+4+5+6+7)/6 = 4.5 > 3.5$. Will every single experiment reveal this ?

Let us toss the two dice each 4 times and average the number of pips. And repeat this 1000 times

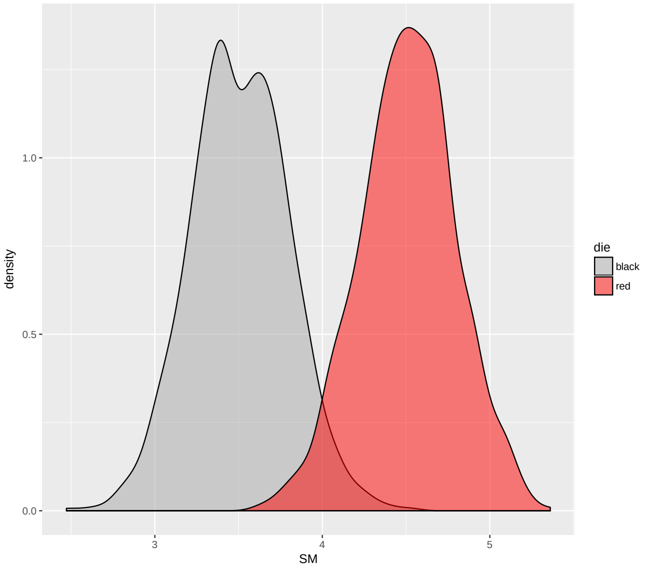
Simulation, N=4

```
## [1] 0.006
```

```
## [1] 0.6
```

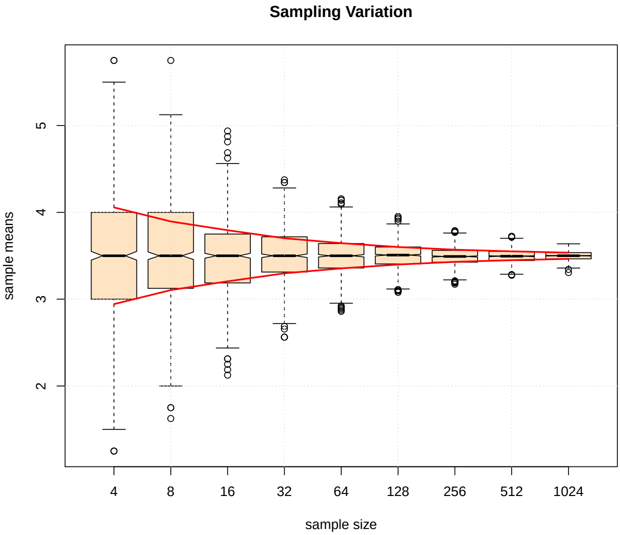


Simulation, N=36

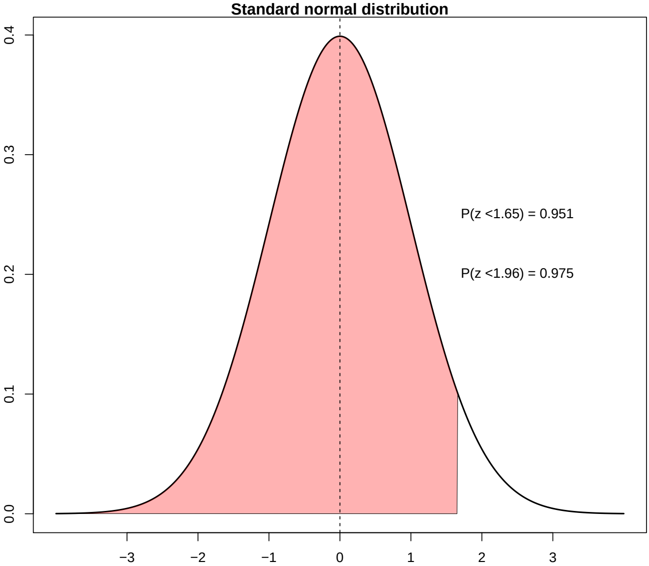


How often were we wrong ? About 0.9% of the time!

Sample Stdev, Scaling



Properties of the Normal distribution



t test

Compare two sets of numbers directly:

```r
N=4
 d1 = sample(1:6,N,T)
 d2 = sample(2:7,N,T)
 #t.test(d1,d2)
 #t.test(d1,d2, paired=TRUE)
 t.test(d1,d2, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  d1 and d2
## t = -2.0494, df = 6, p-value = 0.08631
## alternative hypothesis: true difference in means is not equal
## 95 percent confidence interval:
##  -3.8394488  0.3394488
## sample estimates:
## mean of x mean of y
##      3.00      4.75
```