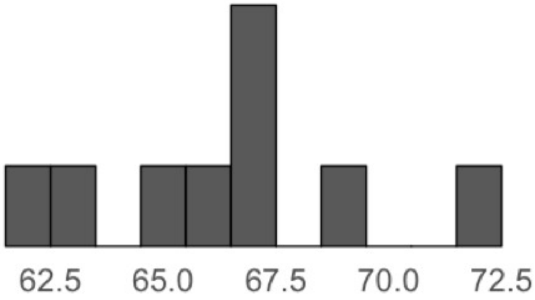
201ab Quantitative methods
L.06: Classical statistics
(with normal -- Z-tests)

Outline

- There is only one test.
- Central limit theorem and normal distribution
- Z-test
- Confidence intervals
- Null hypothesis significance testing
- Power

Our data

(sample of 9 female heights, in inches)



→

A statistic

(arithmetic mean)

mean(x) = 66.44

Inferences we might want to make:

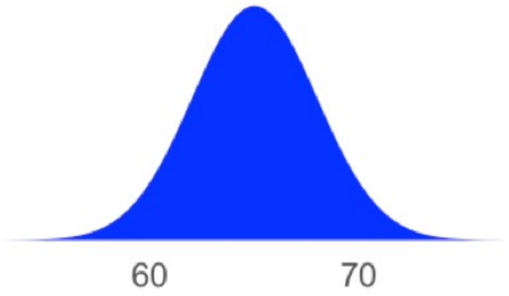
Null Hypothesis Testing: Is this sample likely to have come from a particular known *population* (H0)?

Estimation: What’s the mean of the *population* from which this sample came? What are plausible pop. means?

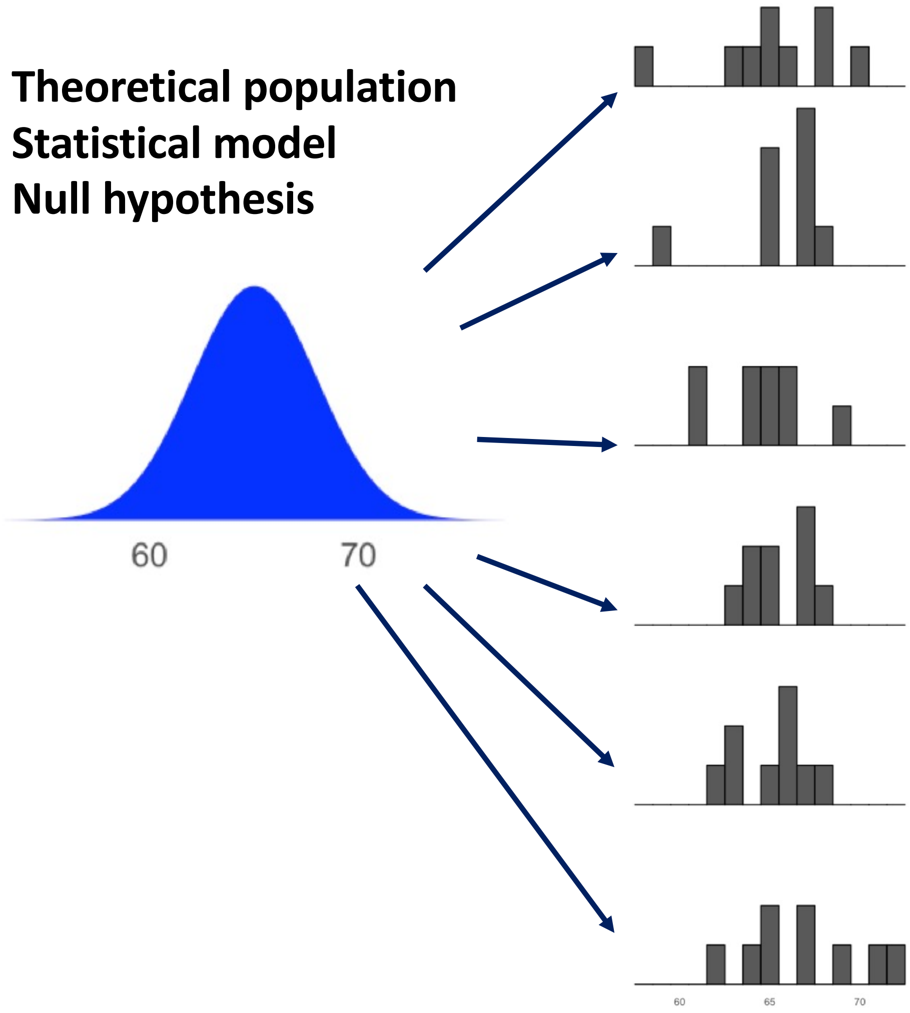
Theoretical population

Statistical model

Null hypothesis

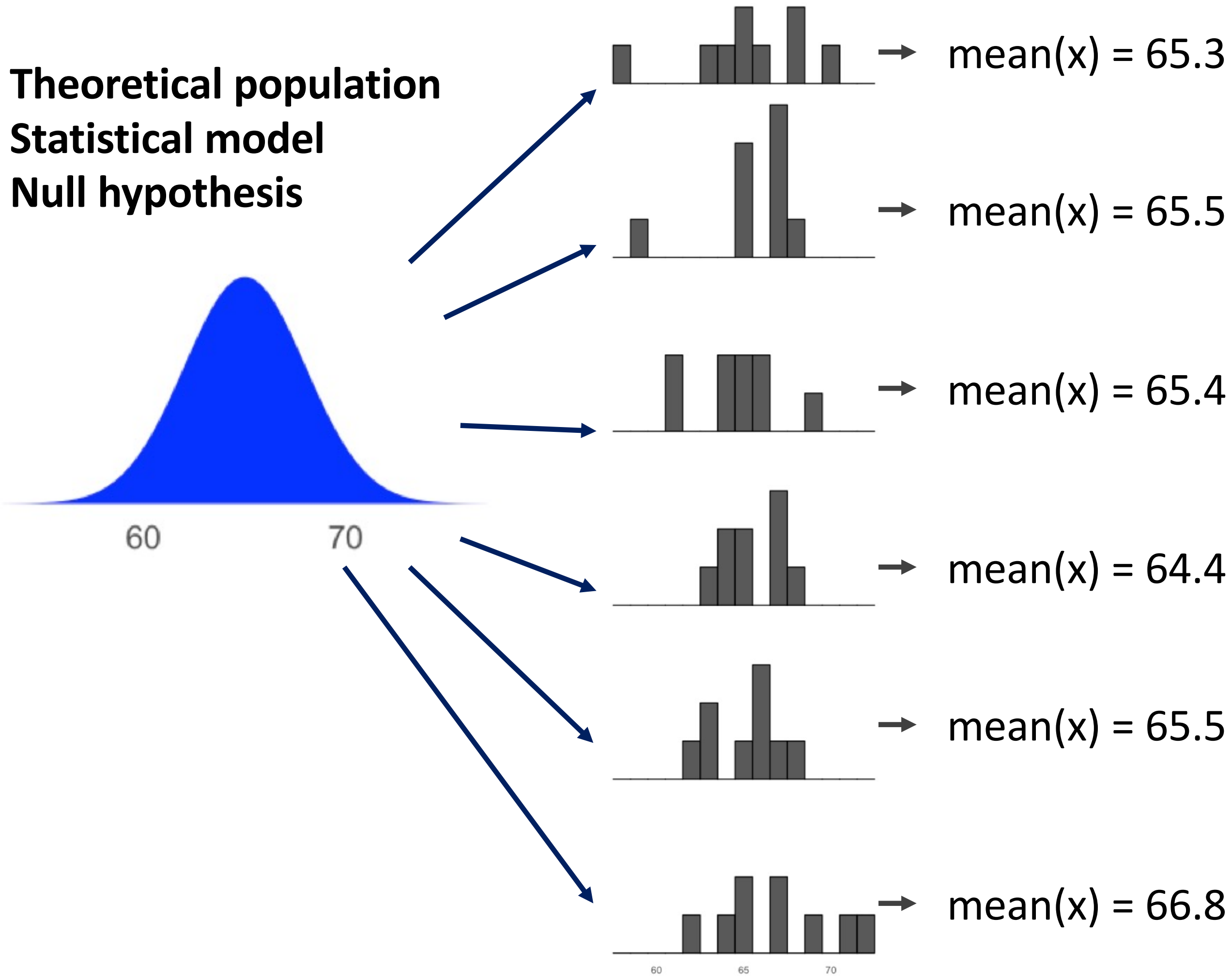


Theoretical population
Statistical model
Null hypothesis



Samples
from the
model
(same size as
our actual sample)

Theoretical population
Statistical model
Null hypothesis



mean(x) = 65.3

mean(x) = 65.5

mean(x) = 65.4

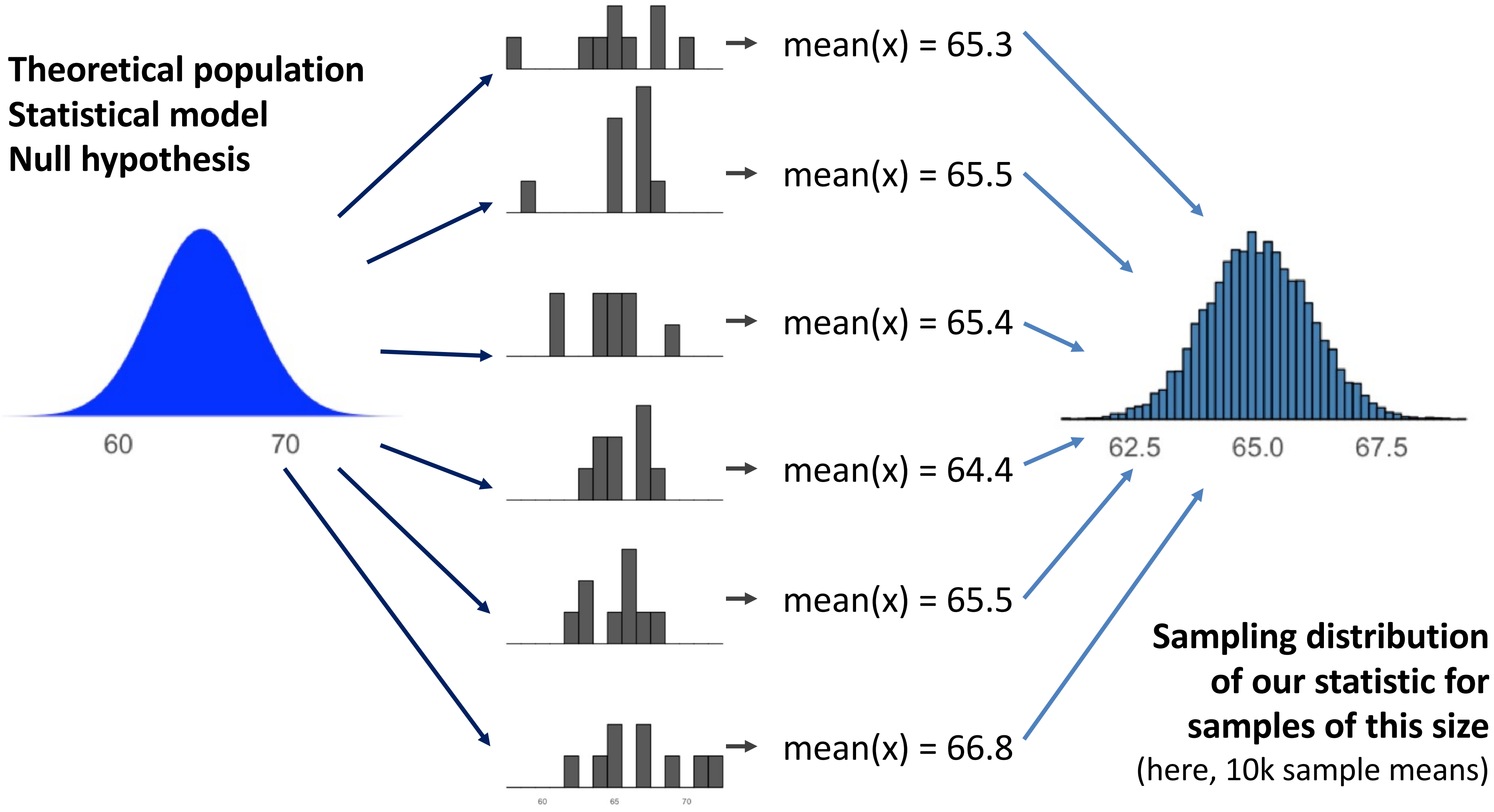
mean(x) = 64.4

mean(x) = 65.5

mean(x) = 66.8

Samples from the model
(same size as our actual sample)

Sample statistics of these samples
(arithmetic means)



Theoretical population
Statistical model
Null hypothesis

mean(x) = 65.3

mean(x) = 65.5

mean(x) = 65.4

mean(x) = 64.4

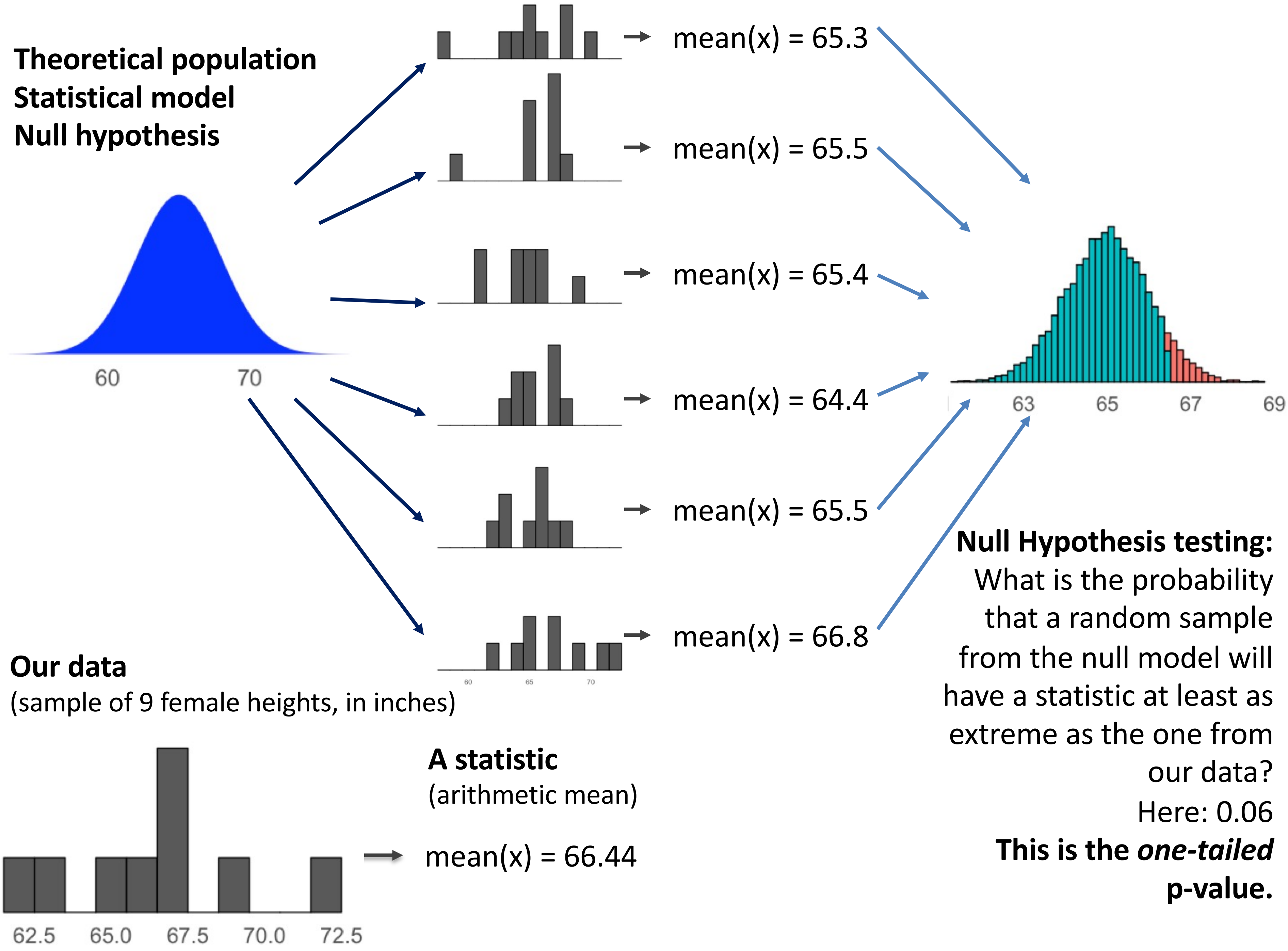
mean(x) = 65.5

mean(x) = 66.8

Samples from the model
(same size as our actual sample)

Sample statistics of these samples
(arithmetic means)

Sampling distribution of our statistic for samples of this size
(here, 10k sample means)



Theoretical population
Statistical model
Null hypothesis

mean(x) = 65.3

mean(x) = 65.5

mean(x) = 65.4

mean(x) = 64.4

mean(x) = 65.5

mean(x) = 66.8

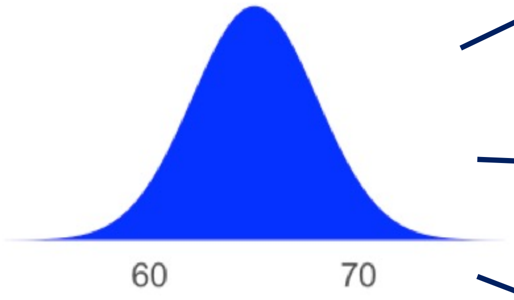
Our data
(sample of 9 female heights, in inches)

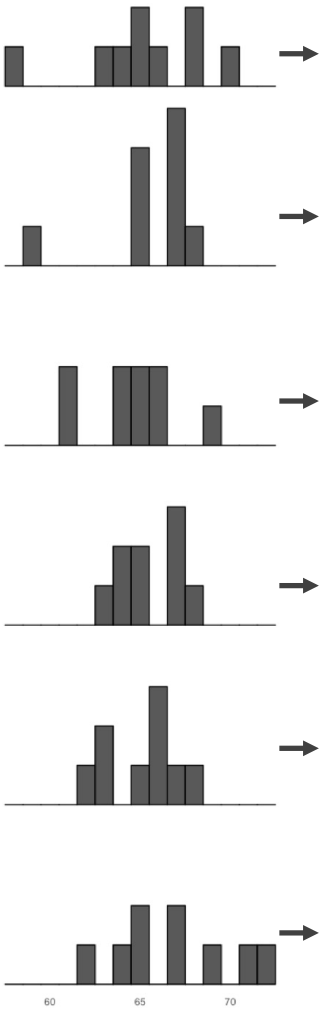
A statistic
(arithmetic mean)

mean(x) = 66.44

Null Hypothesis testing:
What is the probability that a random sample from the null model will have a statistic at least as extreme as the one from our data?
Here: 0.06
This is the *one-tailed* p-value.

Theoretical population
Statistical model
Null hypothesis





sd(x) = 3.34

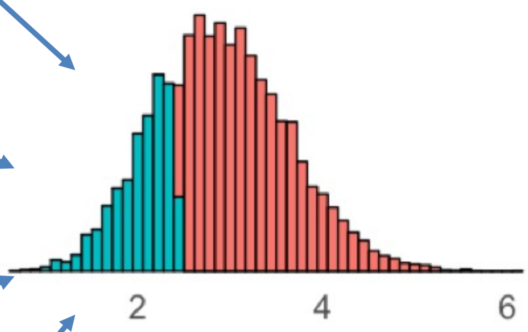
sd(x) = 2.72

sd(x) = 2.66

sd(x) = 1.85

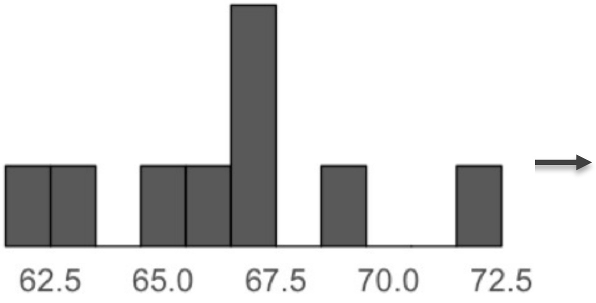
sd(x) = 2.10

sd(x) = 3.43



Null Hypothesis testing:
What is the probability that a random sample from the null model will have a statistic at least as extreme as the one from our data?
Here: 0. 731
This is the *one-tailed* p-value.

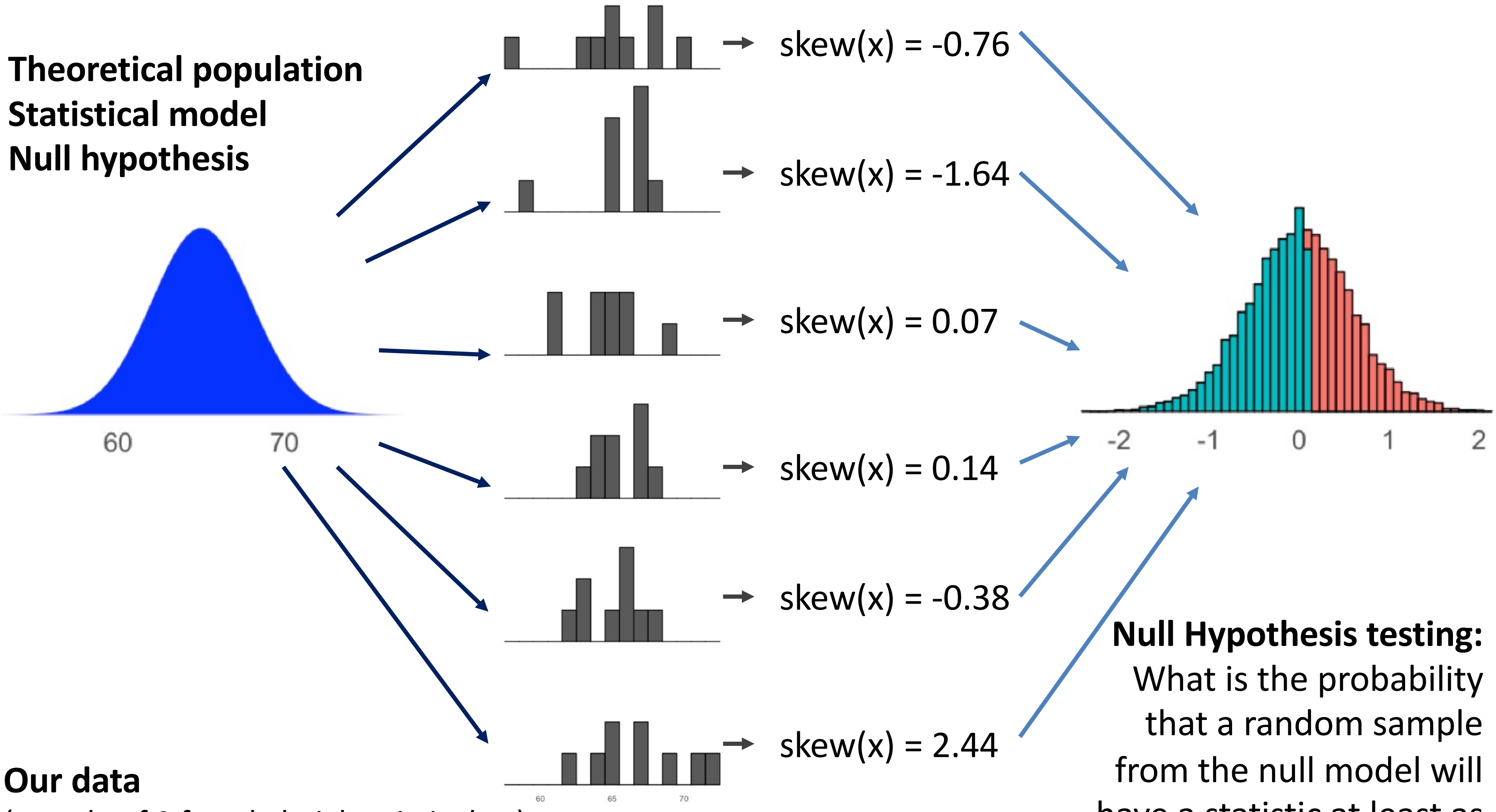
Our data
(sample of 9 female heights, in inches)



A statistic
(standard deviation)

sd(x) = 2.44

Theoretical population
Statistical model
Null hypothesis



skew(x) = -0.76

skew(x) = -1.64

skew(x) = 0.07

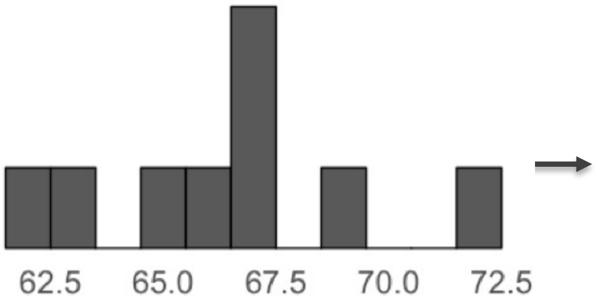
skew(x) = 0.14

skew(x) = -0.38

skew(x) = 2.44

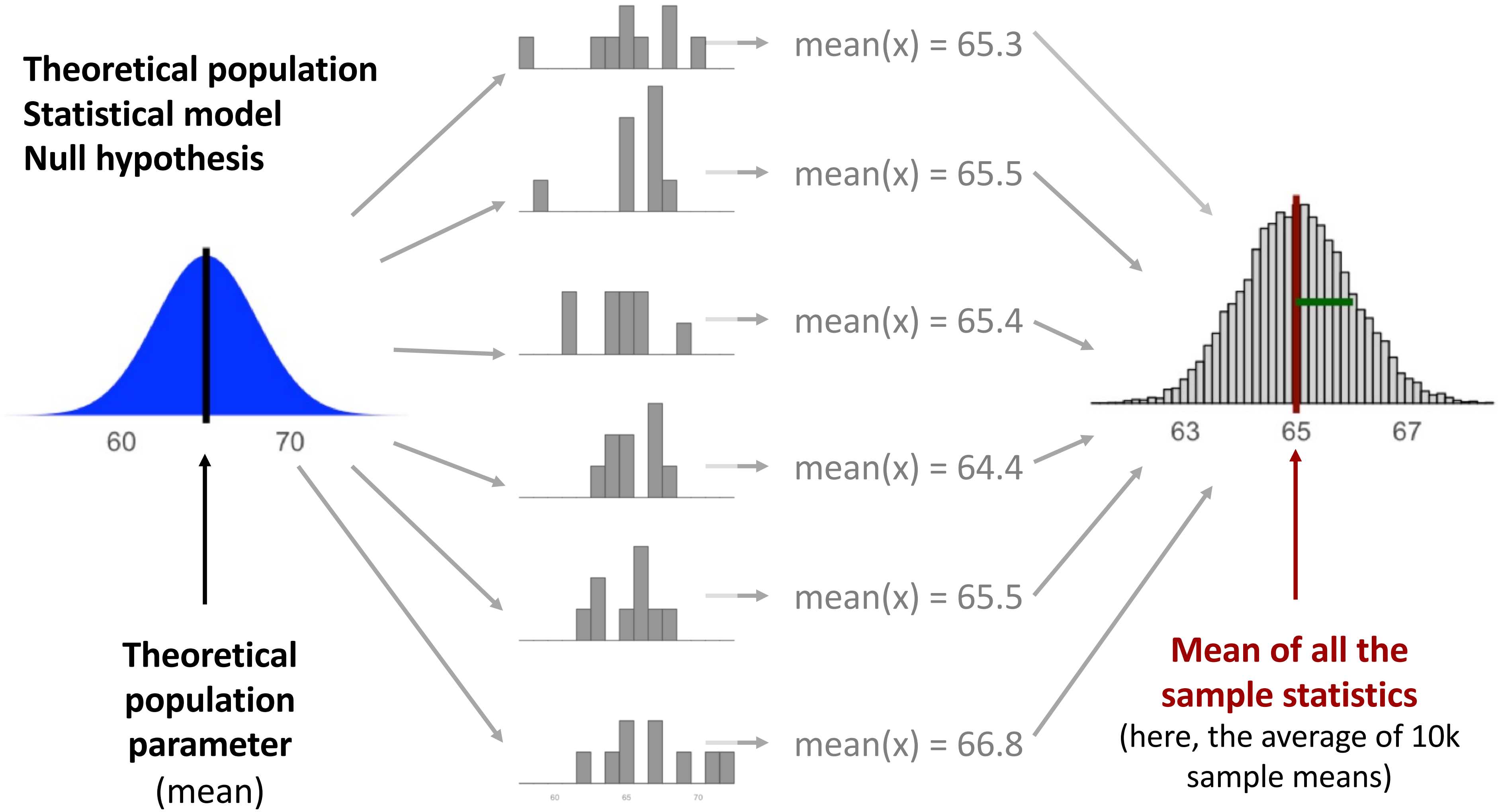
Null Hypothesis testing:
What is the probability that a random sample from the null model will have a statistic at least as extreme as the one from our data?
Here: 0.406
This is the *one-tailed* p-value.

Our data
(sample of 9 female heights, in inches)



A statistic
(skewness)

skew(x) = 0.128



Theoretical population
Statistical model
Null hypothesis

mean(x) = 65.3

mean(x) = 65.5

mean(x) = 65.4

mean(x) = 64.4

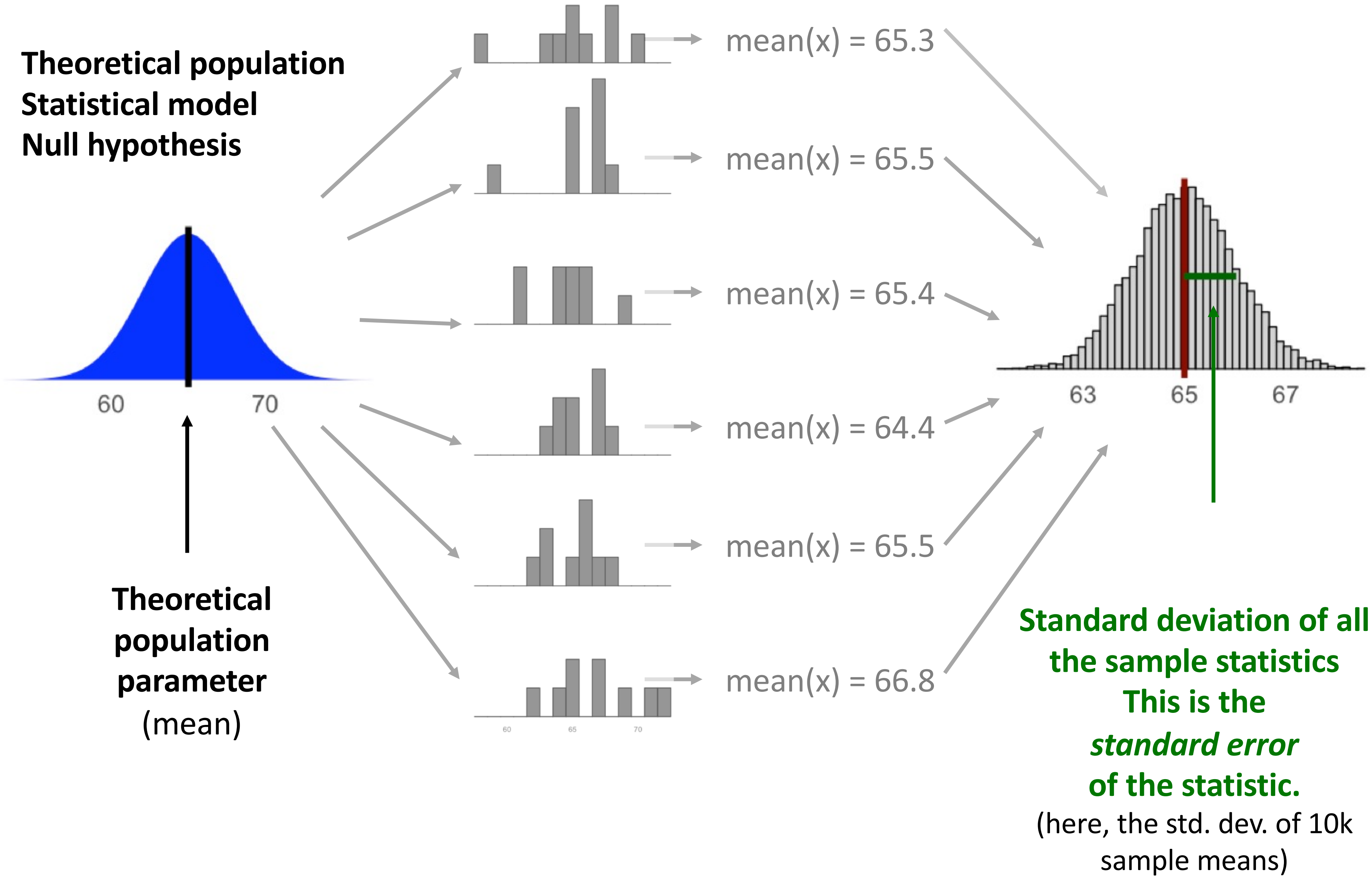
mean(x) = 65.5

mean(x) = 66.8

Theoretical population parameter
(mean)

Mean of all the sample statistics
(here, the average of 10k sample means)

If the average sample statistic has the same value as the population parameter, it is an *unbiased* estimator for that parameter.

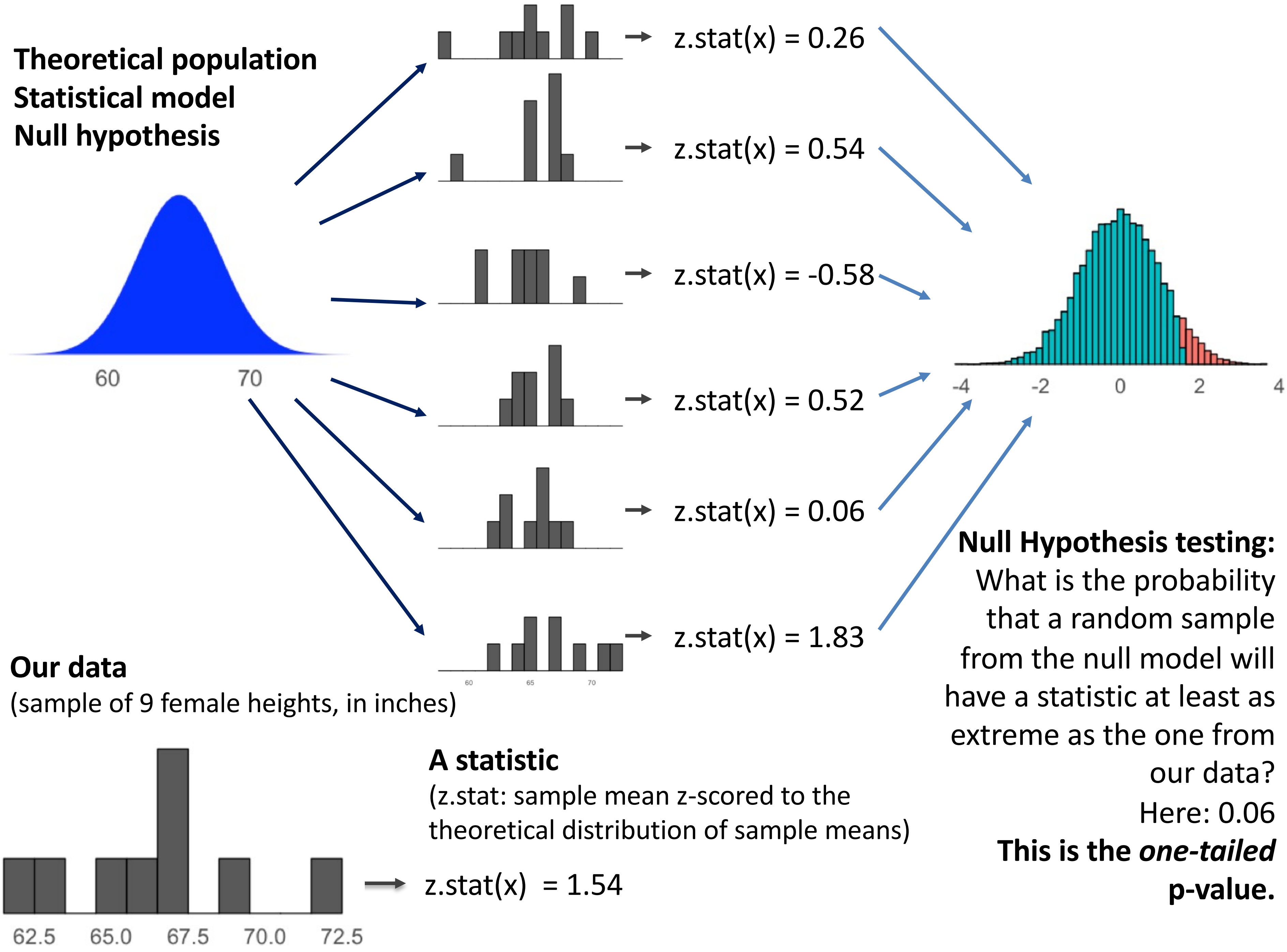


Theoretical population
Statistical model
Null hypothesis

Theoretical population parameter (mean)

Standard deviation of all the sample statistics
This is the ***standard error*** **of the statistic.**
(here, the std. dev. of 10k sample means)

We can calculate a z-score for a given sample statistic by figuring out how many standard errors away it is from the average sample statistic.



Theoretical population
Statistical model
Null hypothesis

z.stat(x) = 0.26

z.stat(x) = 0.54

z.stat(x) = -0.58

z.stat(x) = 0.52

z.stat(x) = 0.06

z.stat(x) = 1.83

Our data
(sample of 9 female heights, in inches)

A statistic
(z.stat: sample mean z-scored to the theoretical distribution of sample means)

z.stat(x) = 1.54

Null Hypothesis testing:
What is the probability that a random sample from the null model will have a statistic at least as extreme as the one from our data?
Here: 0.06
This is the *one-tailed* p-value.

There is only one test.

- There are many significance tests you may have heard of:
  - Z-test, t-test, F-test, X^2-test, etc.
  - These are all named after the statistic they use
- They all follow the same logic:
  - Compare the sample statistic you have to the distribution of sample statistics expected from the null hypothesis.
  - These specific tests are popular because we can analytically derive the sampling distribution of their statistic, and many questions can be posed such that the answer boils down to that statistic.
- The key things to worry about are:
  - what does the statistic measure?
  - what is the null hypothesis?

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals
- Null hypothesis significance testing
- Power

How a generalized statistic is born (z-score)

There is some underlying distribution/population of our data

- **x** ~ P(X)   mean(X) = μ_X, sd(X)=σ_X

Central limit theorem and limit of n -> inf.

- mean(**x**) ~ Normal(...)

Derive from rules of expectation...

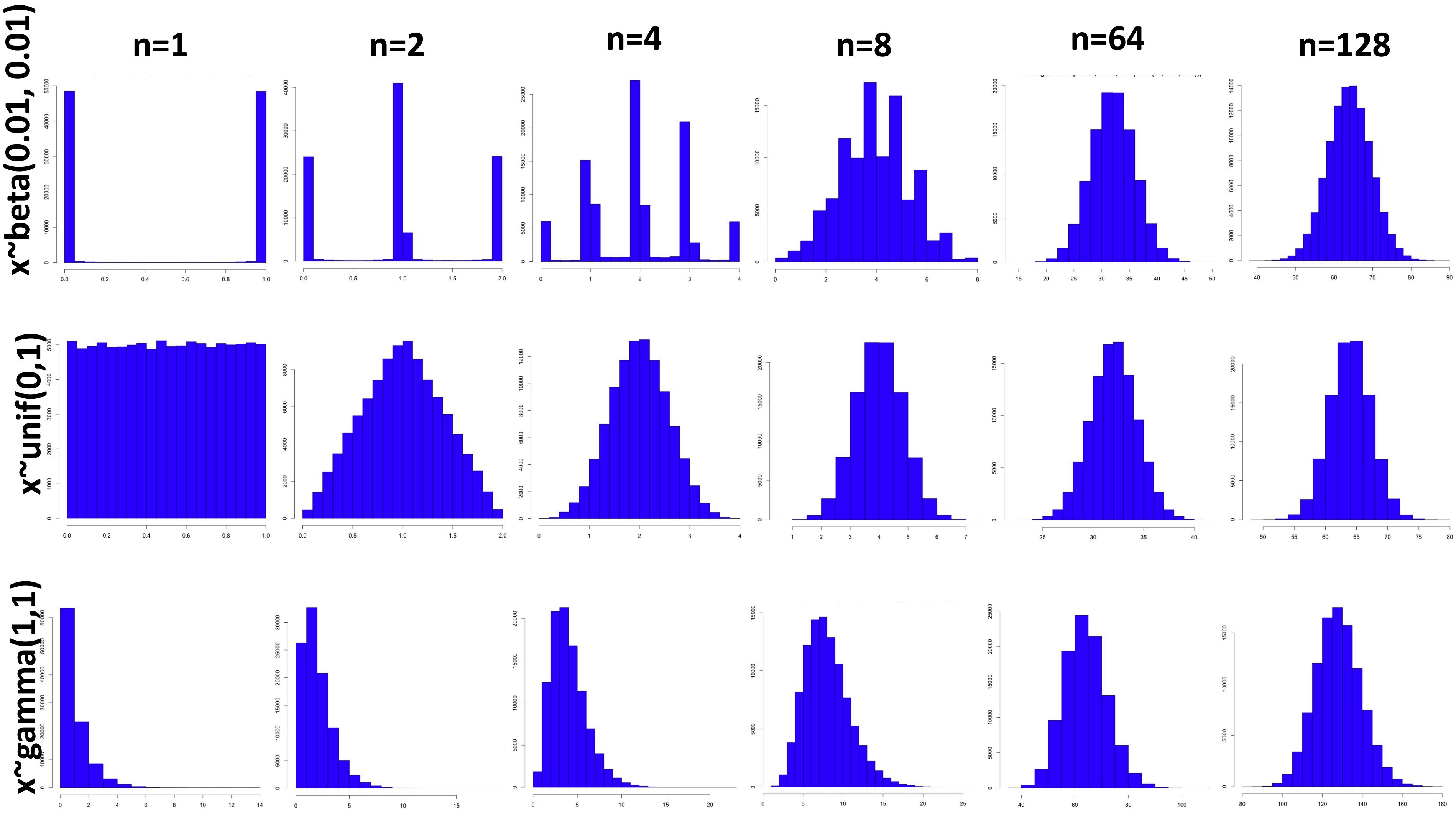
- mean(**x**) ~ Normal(μ_X, σ_X /sqrt(n))

A bit more algebra yields...

- (mean(**x**) – μ_X)/(σ_X /sqrt(n)) ~ Normal(0,1)

- Z(mean(**x**)) ~ Normal(0,1)

Distribution of the sum of *n* iid RVs

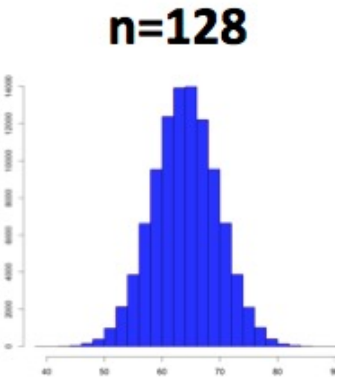


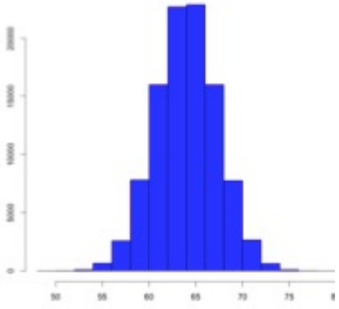
```
replicate(100000, sum(rgamma(n,1,1))
```

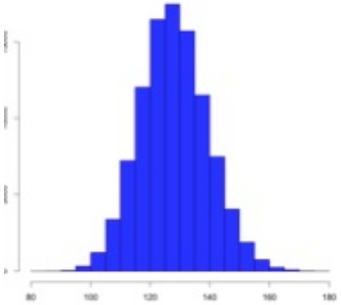
Central limit theorem

- The sum of *n* i.i.d. random variables is Normally distributed if *n* is big enough*

- Many real-world variables can be thought of as the sum of lots of independent and roughly identically distributed, contributing factors, so we often treat our measures as having a Normal distribution, but this should be verified.

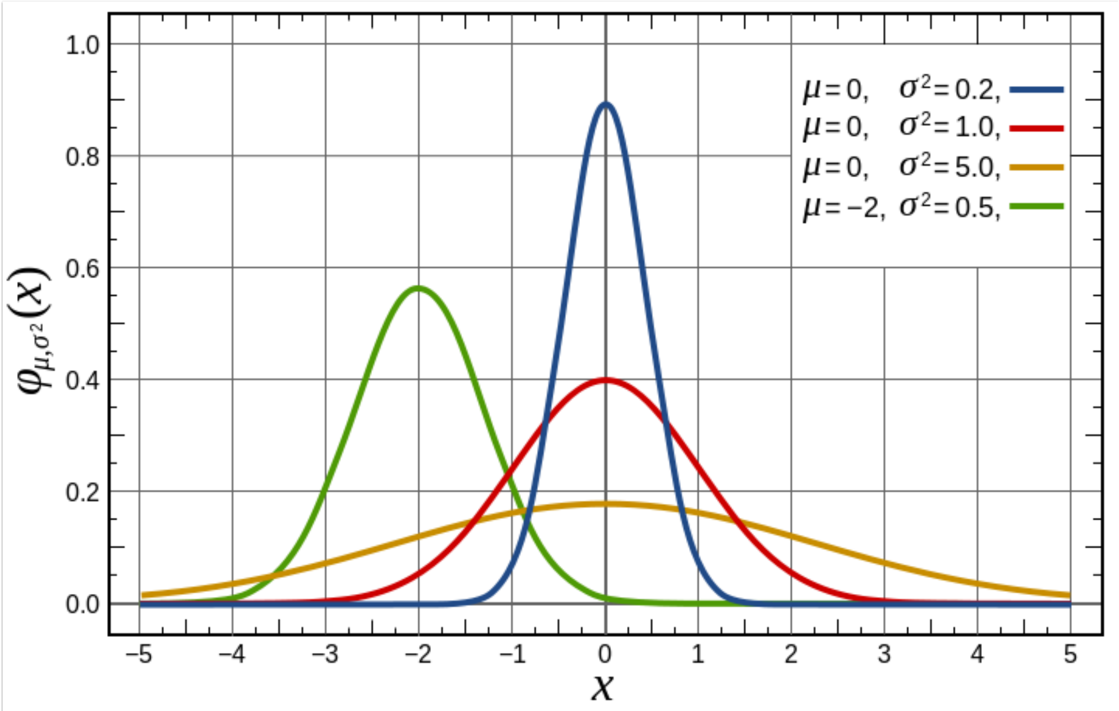






Normal Distribution

It has two parameters:
“location” (mean; mu)
“scale” (sd or var)



In R for a Normal distribution with mean M and sd S

Probability density at x = dnorm(x, M, S)
Cumulative probability at x = pnorm(x, M, S)
Quantile function for p = qnorm(p, M, S)
n Random samples = rnorm(n, M, S)

Sampling dist. of a data point

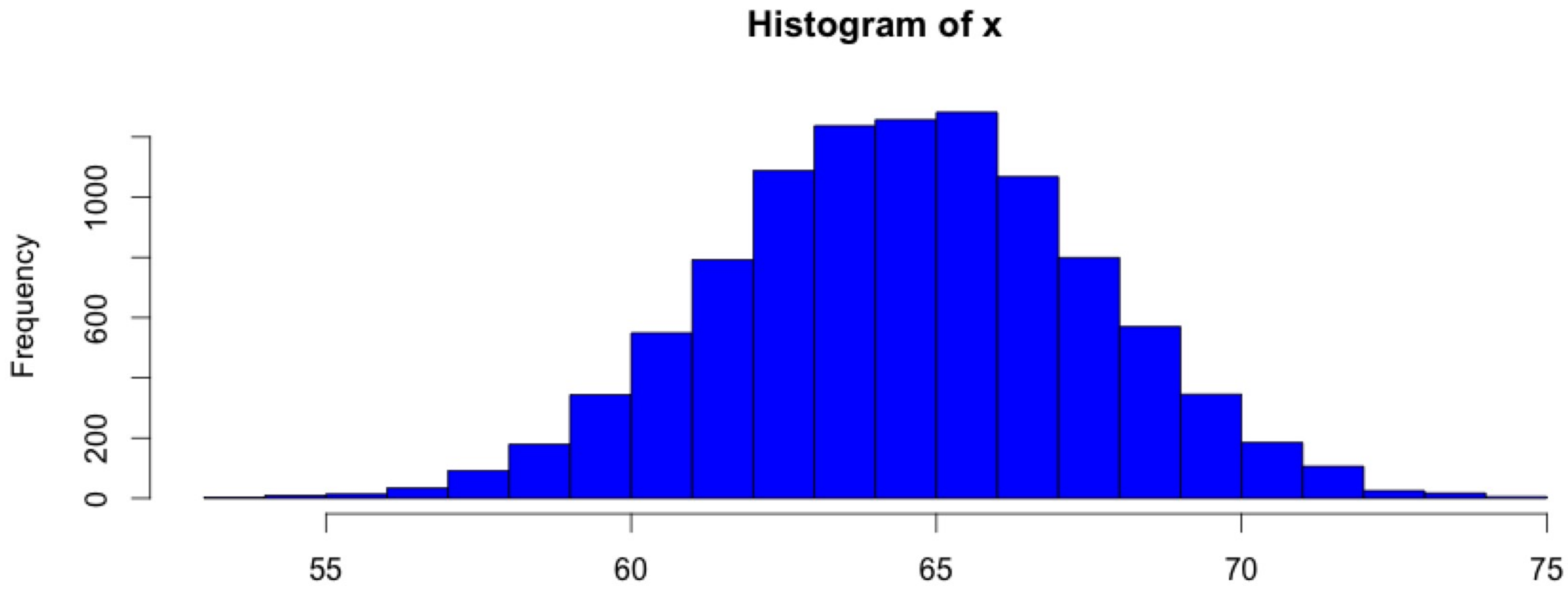
Random samples of CA female heights (Normal, mean=65”, sd=3”)

```
round(rnorm(1,65,3))          →          [1] 67
round(rnorm(1,65,3))          →          [1] 63
```

```
x = round(rnorm(100000,65,3))
```

```
[100000] 61 64 62 64 64 64 70 62 63 59 68 70 66 68 65 64 63 64 65 65 …
```

100000 samples of CA female heights (Normal, mean=65”, sd=3”)



Sampling dist. of the sample mean

```
x = round(rnorm(25,65,3))
                    [25] 61 64 62 64 64 64 70 62 63 59 68 70 66 68 65 64 63 64 65 65 63 72 64 63 62
```

25 samples of CA female heights (Normal, mean=65”, sd=3”)

```
mean(x)                        →                   [1] 64.6
```

Mean of those 25 samples. One possible sample mean.

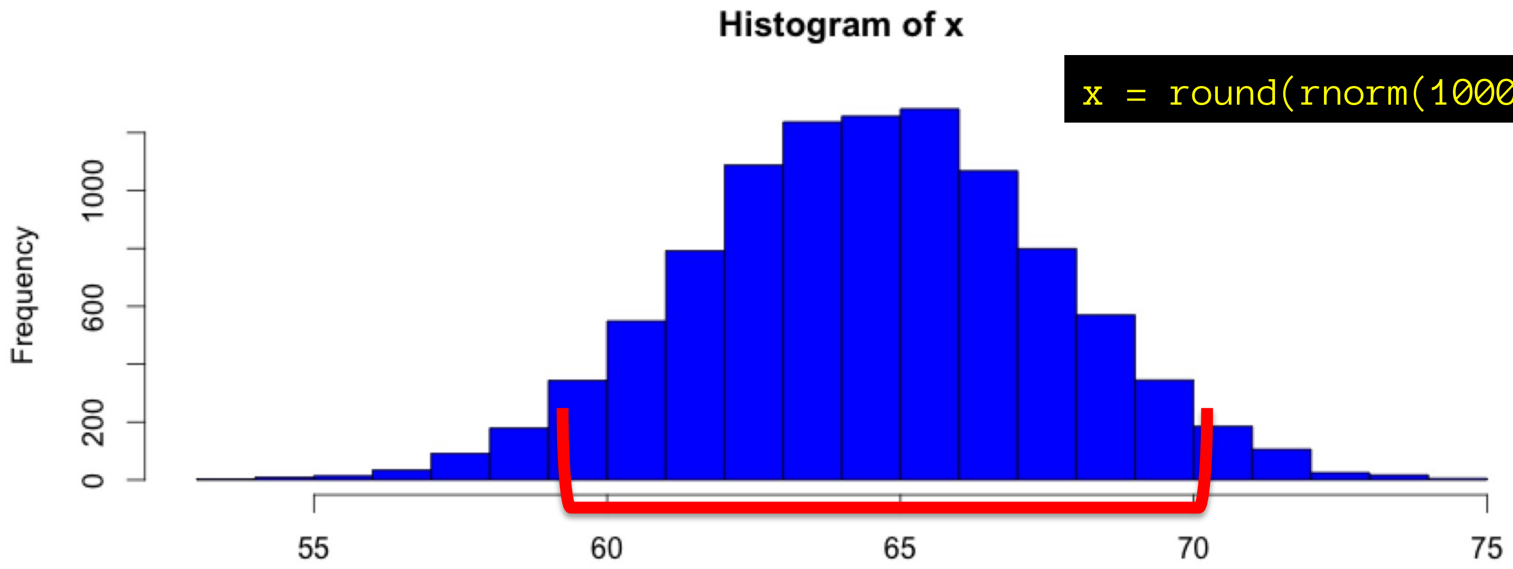
```
mean(round(rnorm(25,65,3)))                        [1] 65.6
mean(round(rnorm(25,65,3)))    →                   [1] 65.6
mean(round(rnorm(25,65,3)))    →                  [1] 65.08
mean(round(rnorm(25,65,3)))    →                   [1] 65.4
```

More sample means of 25 CA female heights.

```
replicate(2, mean(round(rnorm(25,65,3))))    [2] 65.04 65.80
replicate(10000, mean(round(rnorm(25,65,3))))
             [10000] 64.40 64.44 65.20 65.36 65.44 64.56 64.68 ...
```

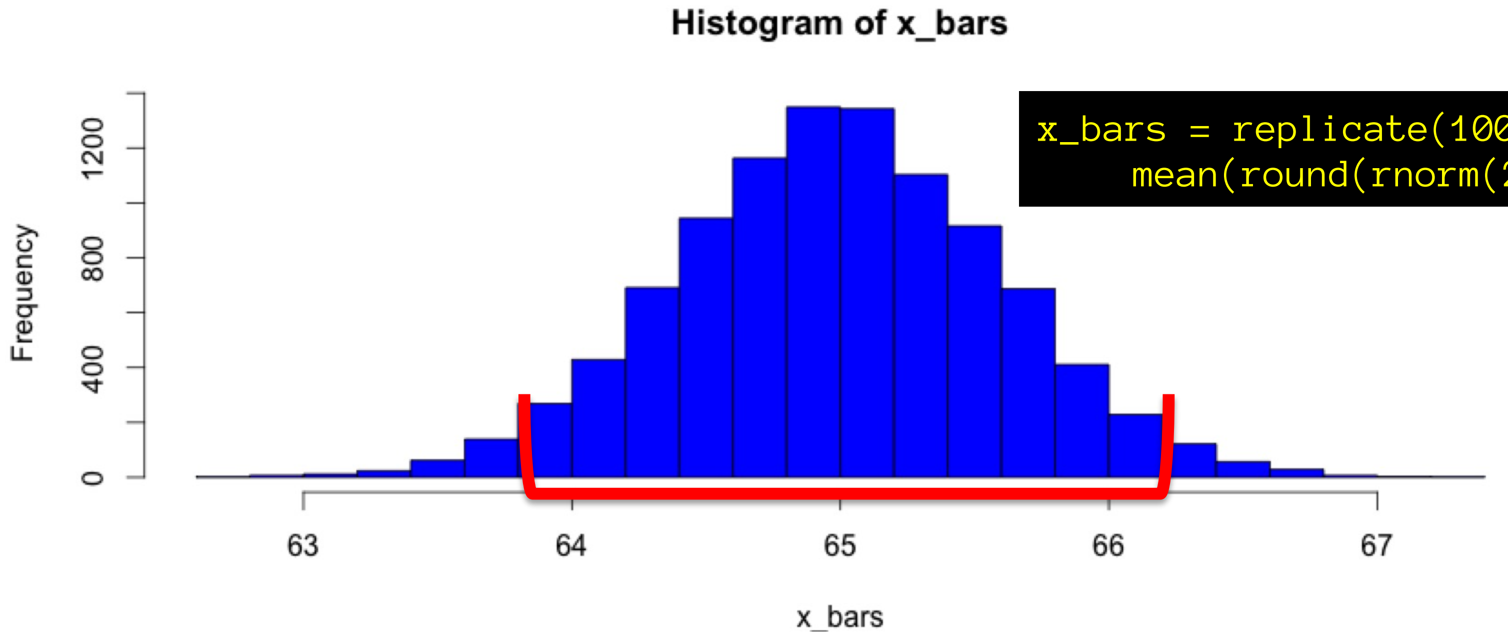
We generate many sample means at the same time with replicate.

Sampling dist. of sample mean



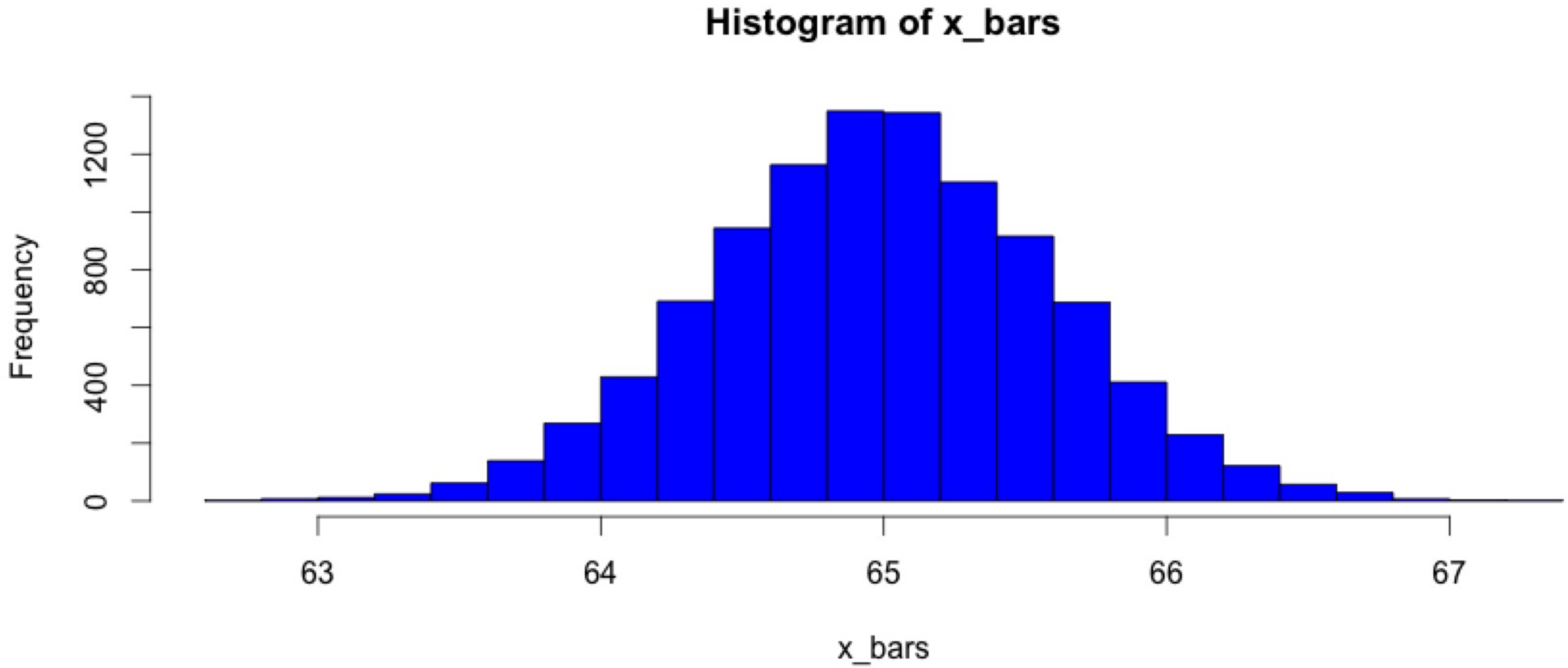
```
x = round(rnorm(100000,65,3))
```

Sampling distribution of sample mean has smaller sd by 1/sqrt(n)



```
x_bars = replicate(10000,
    mean(round(rnorm(25,65,3))))
```

Sampling dist. of the sample mean



Histogram of many sample means of n=25 samples of female heights Normal(mean=65, sd=3).
Follows a Normal(mean=65, sd=3/sqrt(25))

This is the sampling distribution of this sample mean.

Sampling dist. of the sample mean

$$\{x_i,\ldots,x_n\}\underset{iid}{\sim}P(X)$$

We take n samples from some population represented by a probability distribution P(X)

$$Mean[X]=\mu_X$$

The population mean is mu_x

$$Variance[X]=\sigma_X^2$$

The population variance is sigma_x^2

$$\bar{x}_{(n)}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

The sample mean is the sum of those samples divided by their count.

$$Mean[\bar{x}_{(n)}]=\mu_X$$

$$Variance[\bar{x}_{(n)}]=\sigma_X^2\,/\,n$$

$$\bar{x}_{(n)}\sim Normal\left(\mu_X,\sigma_X\,/\,\sqrt{n}\right)$$

Altogether, the sample mean will be normally distributed (if n is large enough – CLT), around the population mean, with a standard deviation that decreases with sqrt(n).

Sampling dist. of sample mean

```
x = round(rnorm(100000,65,3))
[100000] 61 64 62 64 64 64 70 62 63 59 68 70 66 68 65 64 63 64 65 65 …
```

```
sd(x)
[1] 3.024
```

Oh good. Our samples have the SD we told them to have.

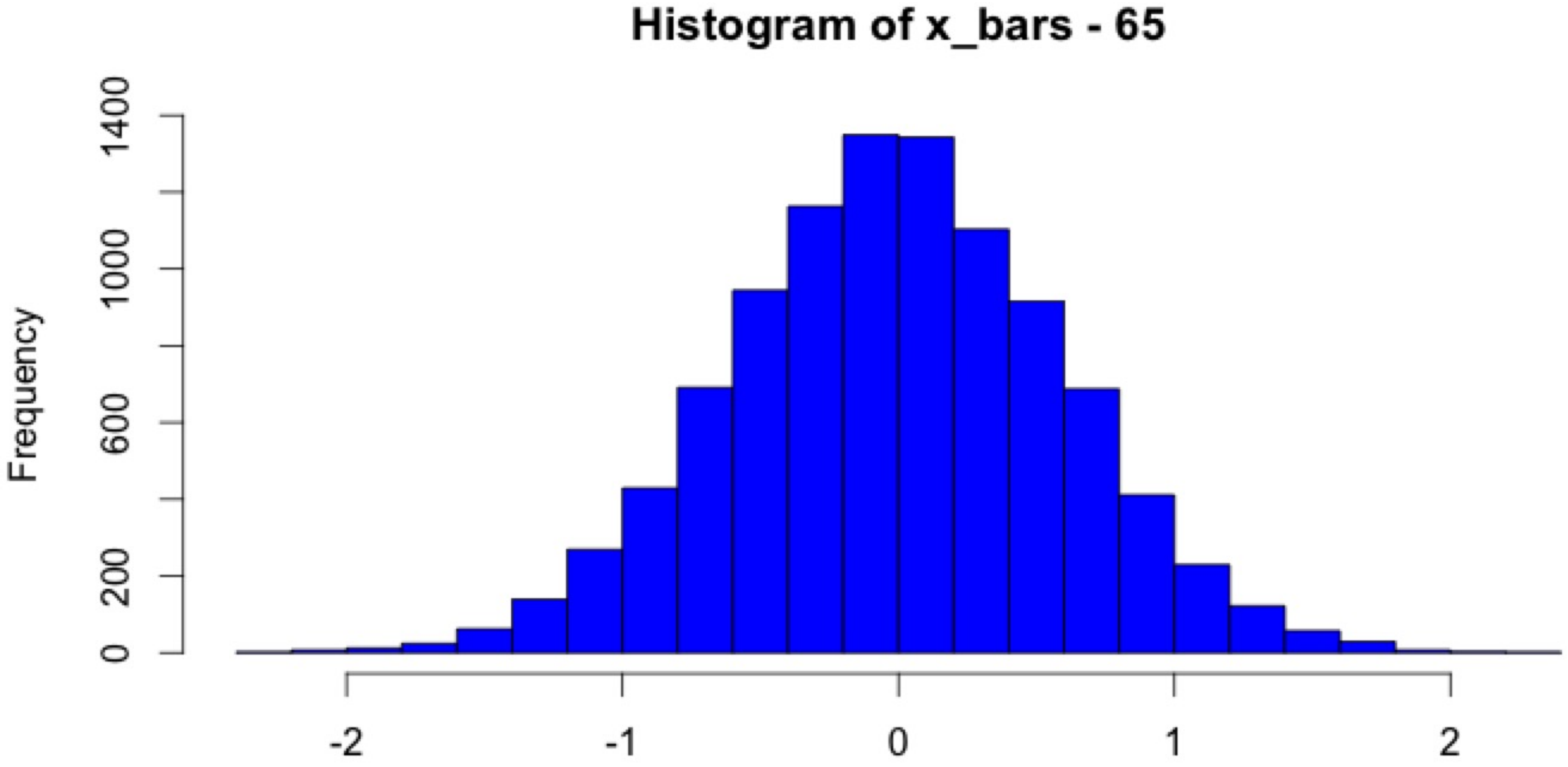
```
x_bars = replicate(10000, mean(round(rnorm(25,65,3))))
[10000] 64.40 64.44 65.20 65.36 65.44 64.56 64.68 ...
```

```
sd(x_bars)
[1] 0.6007
```

And our sample means have the sd they should have according to math.

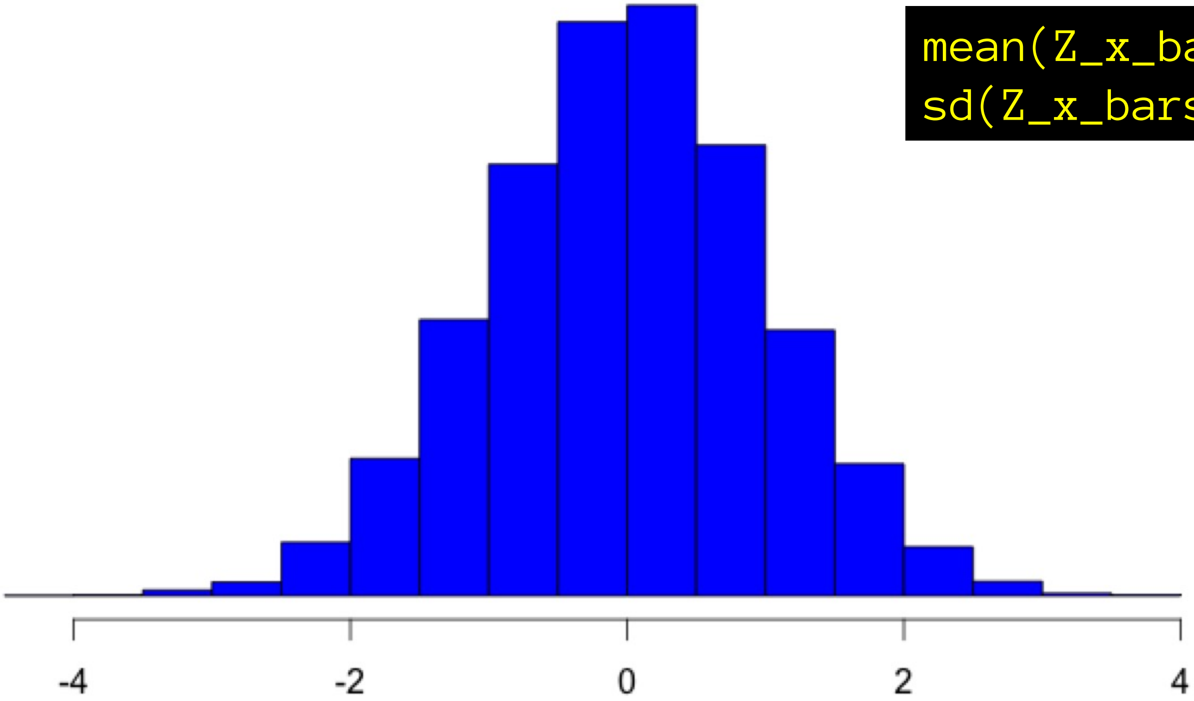
$$\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 3/5 = 0.6$$

Samp. dist. of error of sample mean



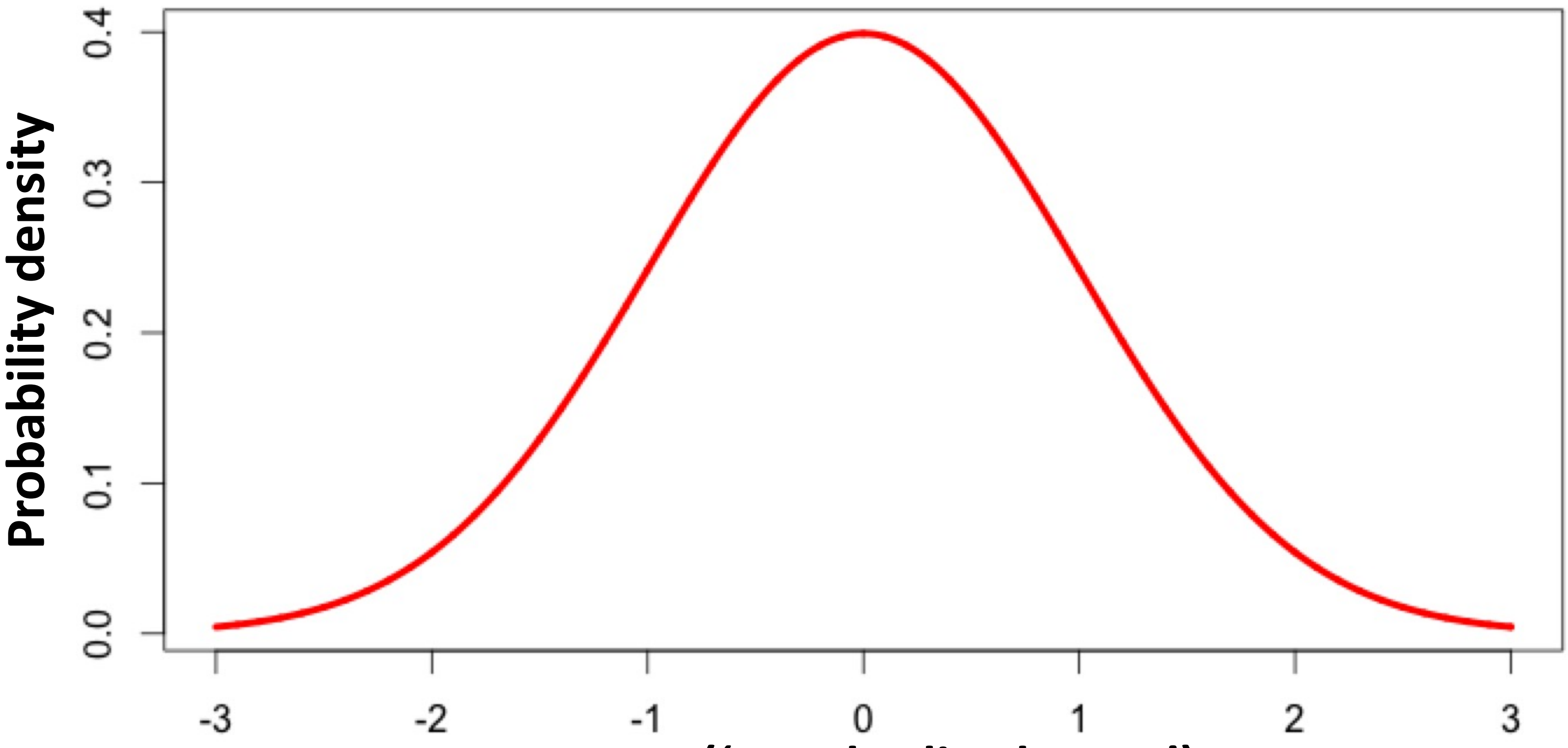
Z_x_bar = (x_bar– pop mean)/sem

```
n=25
sdX = 3
muX = 65
sem = sdX/sqrt(n)
x_bars = replicate(10000, mean(rnorm(n,muX,sdX)))
Z_x_bars = (x_bars - muX)/sem
```



```
mean(Z_x_bars)          [1] -0.004
sd(Z_x_bars)            [1]  1.005
```

The Z (standard normal) distribution



Z score (‘standardized score’)

Distance from the mean in units of standard deviation

Right now: distance between sample mean and population mean in units of standard error of the mean.

Z scores

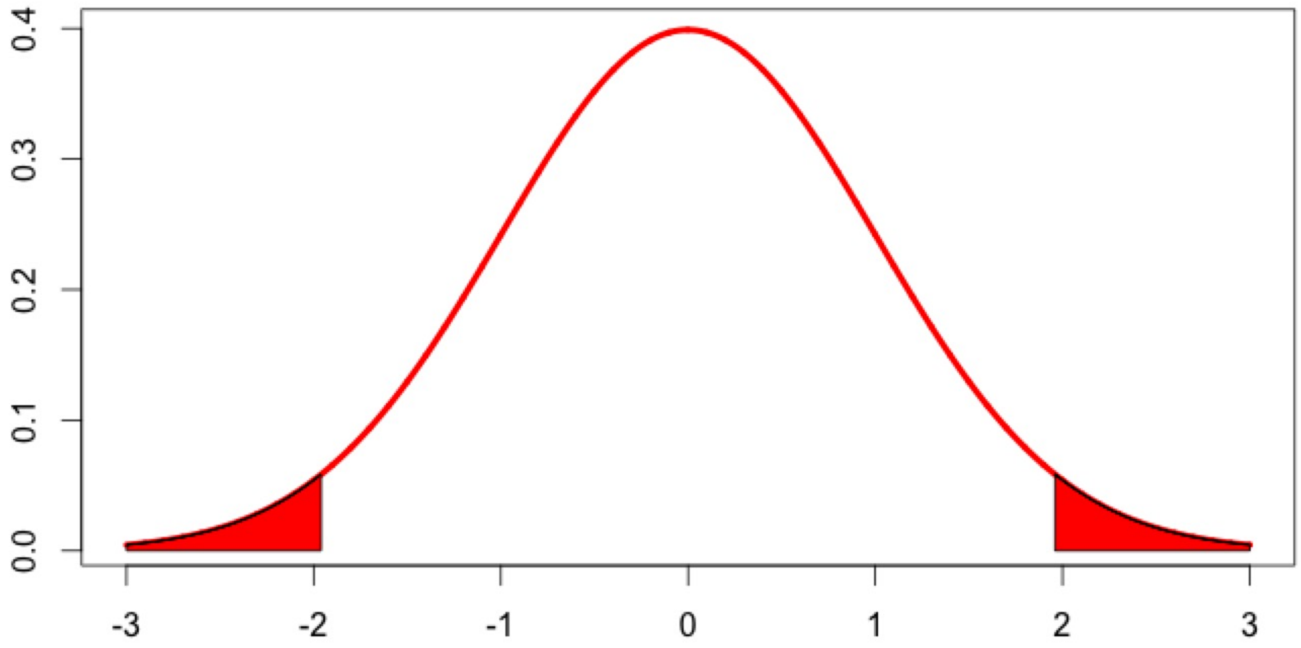
- What is the probability that our sample mean will have a Z-score > 1.96 or < -1.96?
  (i.e. will be more than 1.96 standard errors away from the population mean?)

```
pnorm(-1.96) + (1-pnorm(1.96))
```

```
[1] 0.05
```

```
2*pnorm(-1.96)
```

Equivalent because distribution is symmetric around 0.



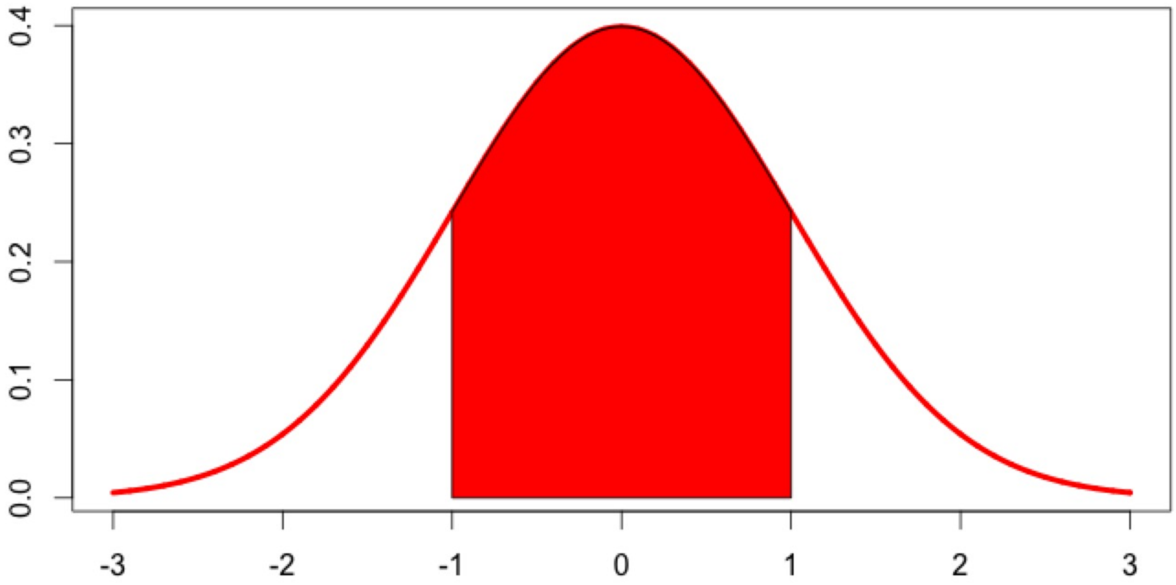
Z scores

- What is the ‘critical’ absolute Z value such that the Z-score of our sample mean will have an absolute value less than that with probability 68.27%?

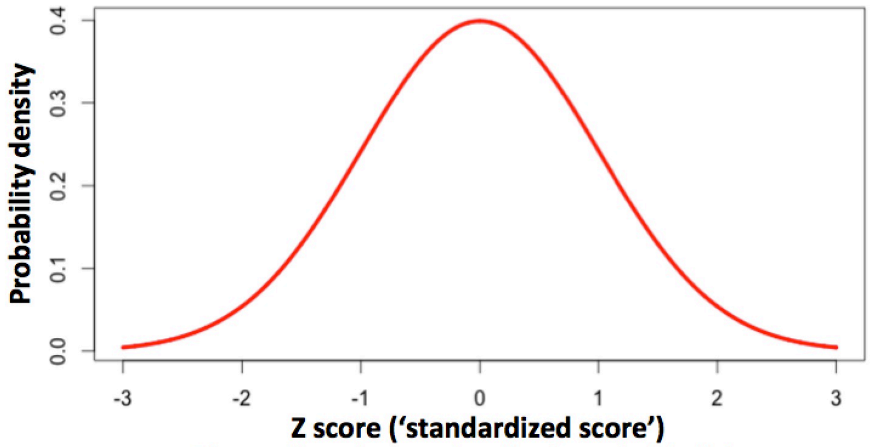
```
(1-0.6827)/2                                          [1] 0.15865
```

That’s how much probability should be ‘left over’ in either tail.

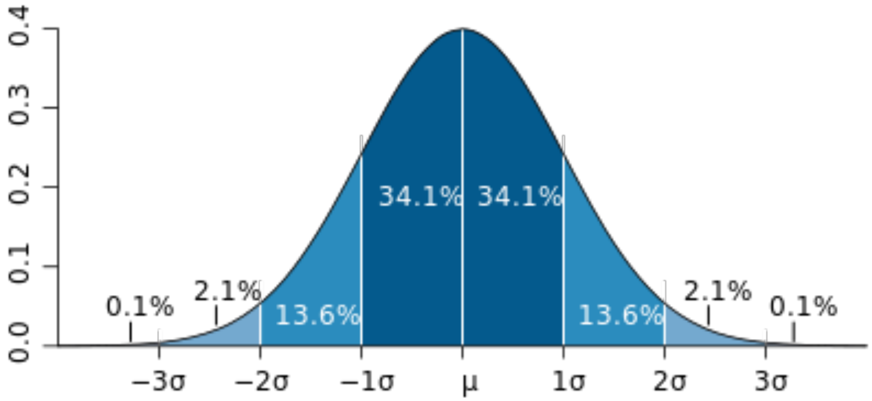
```
qnorm(0.15865)                                             [1] -1
```



Z score distribution – fun facts

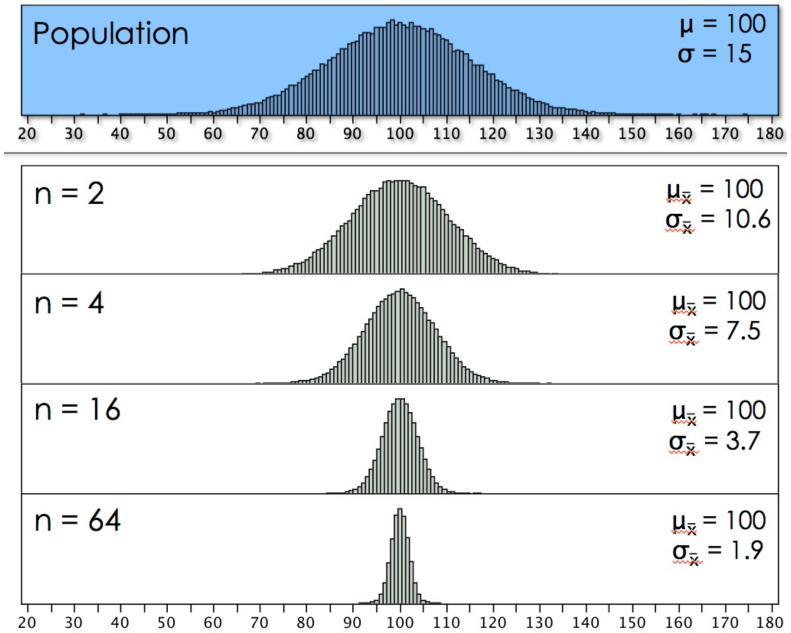


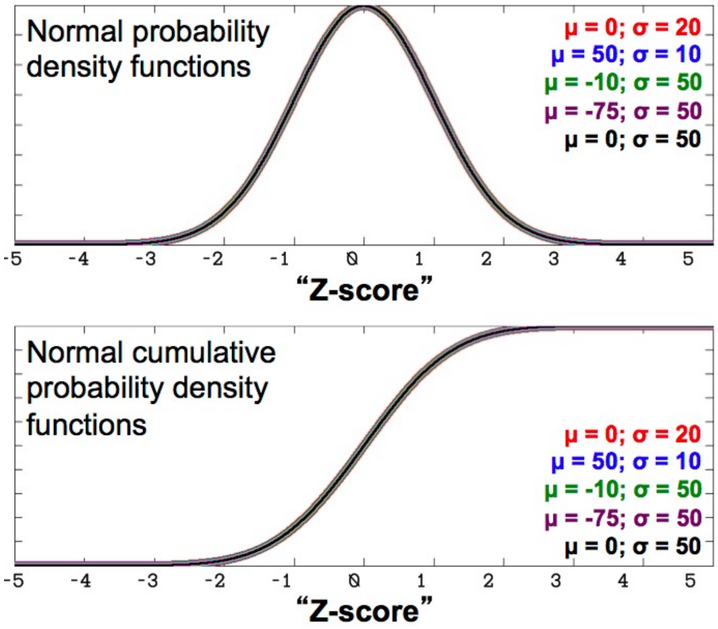
Distance from the mean in units of standard deviation
Right now: distance between sample mean and population mean in units of standard error of the mean.



- z_x = (x-mu)/sd [***]
  relative to distribution of x!
  e.g., (x_bar – mu_x)/sem
  Distance from the mean in standard deviations.
- P(abs(z) < 1) = 0.68
  i.e. 68% of values are less than 1 s.d. away from mean.
- P(abs(z)>1.96) = 0.05
  i.e. 5% of values are more than 1.96 s.d.s away from mean.
- P(abs(z)>1.64) = 0.1
  i.e. 10% of values are more than 1.64 s.d.s away from mean.

Theoretical distributions





Sampling distribution of the sample mean

$$x_i \underset{iid}{\sim} P(X)$$

$$\bar{x}_{(n)} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x}_{(n)} \sim Normal\left(\mu_{\bar{x}} = \mu_X, \sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}\right)$$

Sampling distribution of a Z-score

$$x_i \sim Normal(\mu_X, \sigma_X)$$

$$z_i = \frac{x_i - \mu_X}{\sigma_X}$$

$$z_i \sim Normal(0,1)$$

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals
- Null hypothesis significance testing
- Power

Null hypothesis significance testing.

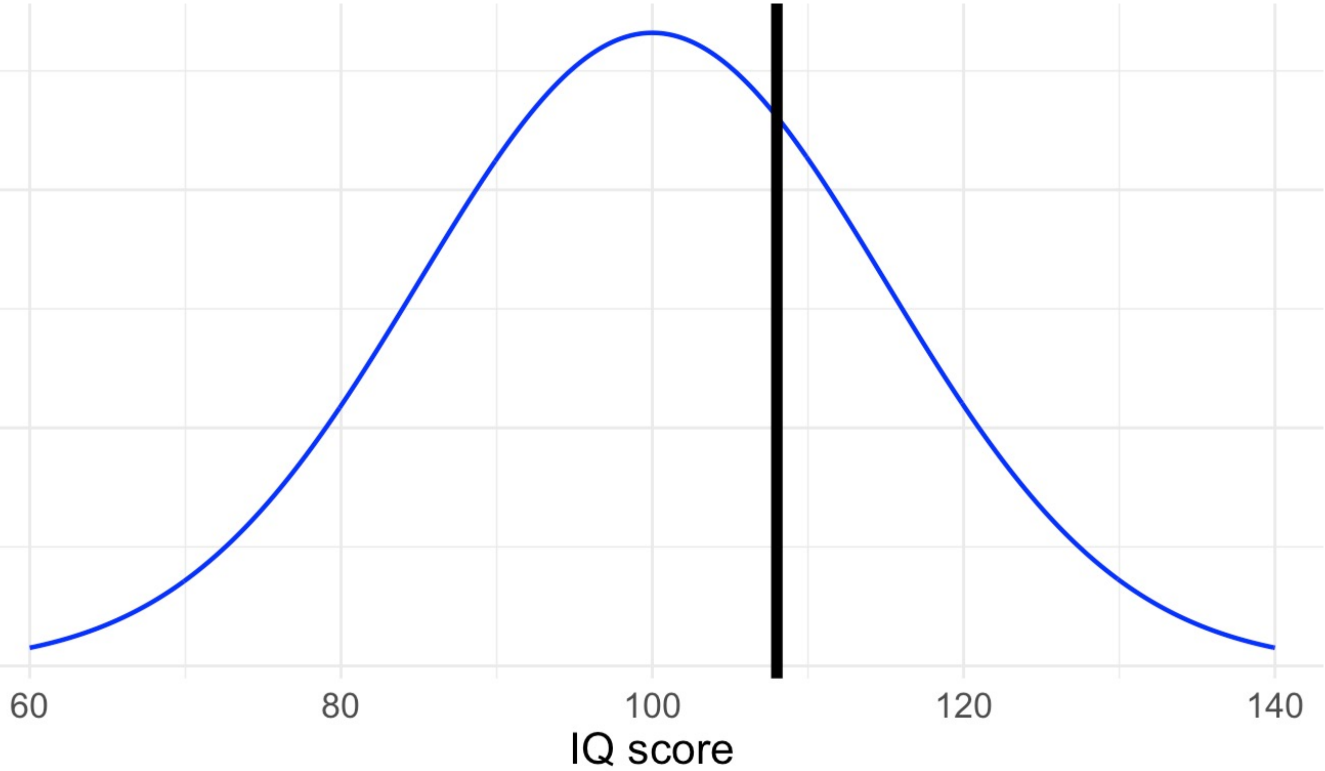
- We have a sample
  - E.g. 16 IQ scores from psych 201
- We have some ‘null’ hypothesis (*statistical model!*)
  - E.g., 201 students are no different from the rest of the population. They are random samples from the overall IQ distribution with mean=100, and sd=15
- We see if the sample is *sufficiently different* from what we expect of samples from the null population, to ‘reject the null’

Null hypothesis significance testing

A (fictional) sample of 16 IQ scores of 201 students.

```
x                                        [16] 120 113 129 113 …
```

We think these folks might have a different *mean* (but same sd) as the normal population.



Sample mean of x compared to null hypothesis distribution of x values

Is this "sufficiently different" to reject the null?

“Sufficiently different”

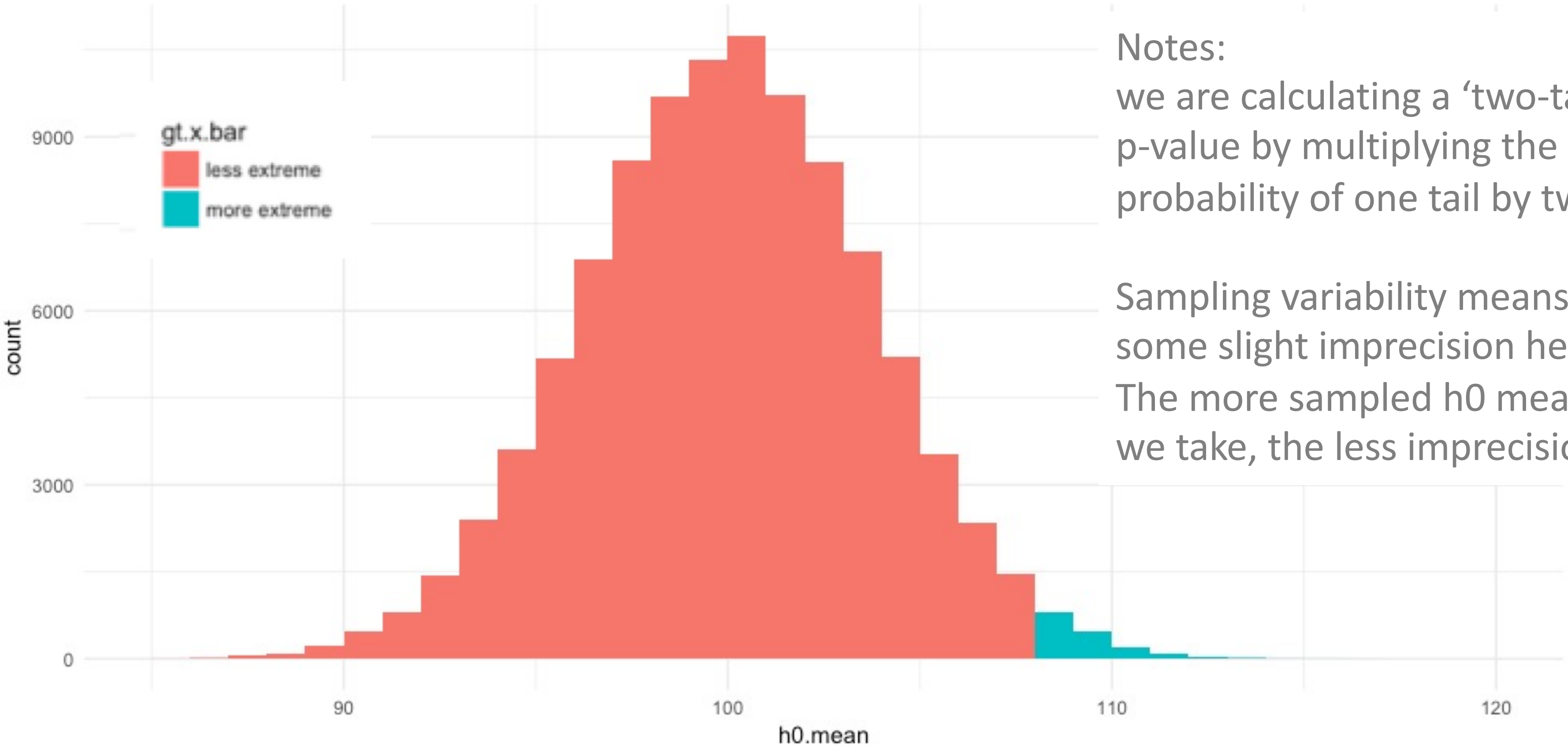
- How likely is an outcome *at least this extreme* to arise if the null hypothesis is true?
  (i.e., in samples from the null model)
- We are going to choose a criterion of 5% (“alpha”):
  If an outcome at least as extreme as this one has more than a 5% chance of arising under the null hypothesis, we deem it not *sufficiently surprising* and we do ***not*** reject the null model. Otherwise we will.
- Why 5%?
  Because Fisher thought it would be ok, then everyone in social science started using it. Physics has much more stringent criteria: 0.000000002 for a “discovery” and 0.0000006 for a cautious announcement. As we will see in the homework, when testing surprising (low base-rate) effects, we can only achieve a reasonable positive predictive value, by adopting a more stringent alpha criterion.

Random sampling for NHST

```
n = 16
x.bar = 108
sample.mean.h0 = function(n){mean(rnorm(n, 100, 15))}
h0.means = replicate(100000, sample.mean.h0(n))
```

```
(cur.p.val = 2*mean(h0.means > x.bar)
```

```
[1] 0.03312
```



Notes:
we are calculating a ‘two-tailed’ p-value by multiplying the probability of one tail by two.

Sampling variability means some slight imprecision here. The more sampled h0 means we take, the less imprecision.

Why don’t we always just do this?

- It’s unconventional.
- It requires some programming and a bit of thought to pick a good statistic to sample in more complex tests.

- With modern computers we can. (randomization: 201b)

- Back in the day: no (machine) computers.

  Strategy: define a simple mathematical transformation that yields one invariant *sampling distribution* for a family of null models. Then the *hard* CDF calculations need to be done only once, and can apply for everyone.

  Thus many null hypothesis tests reduce to one of a few common test-statistics: z, t, F, X2, etc..

Null hypothesis significance testing

- Use the *Z score of the sample mean, relative to the null hypothesis sampling distribution of the sample mean.*

```
x                                            [16] 120 113 129 113  …
```

```
mean(x)                                                    [1]108
```

```
z = (mean(x) - 100)/(15/sqrt(length(x)))             [1] 2.133
```

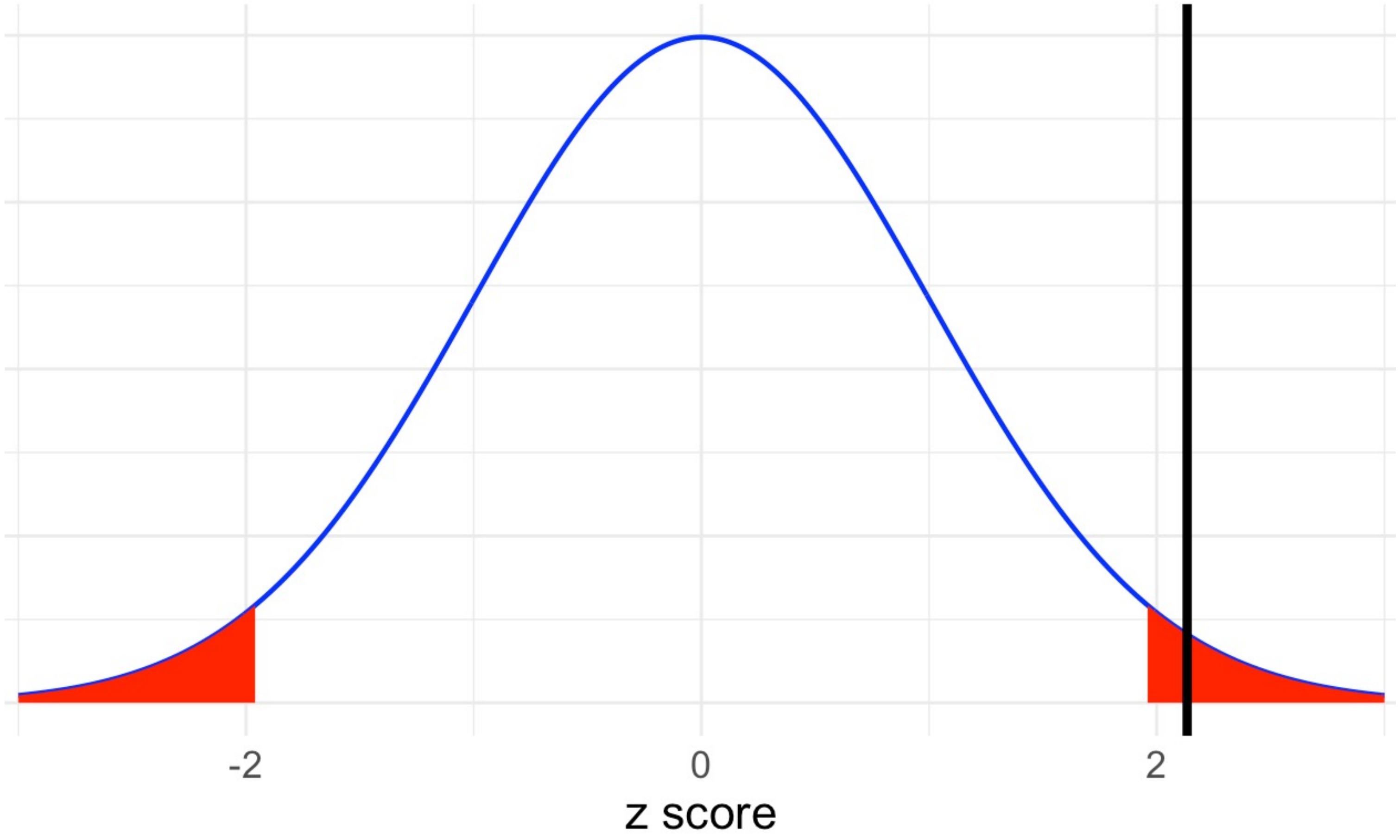
- How big does z have to be to exceed a 5% criterion?

```
Z_crit = qnorm(0.05/2)                                [1] -1.96
```

- Absolute value > 1.96.
- Here it is bigger, so we reject the null at alpha=5%

Standard normal (Z) dist. for NHST

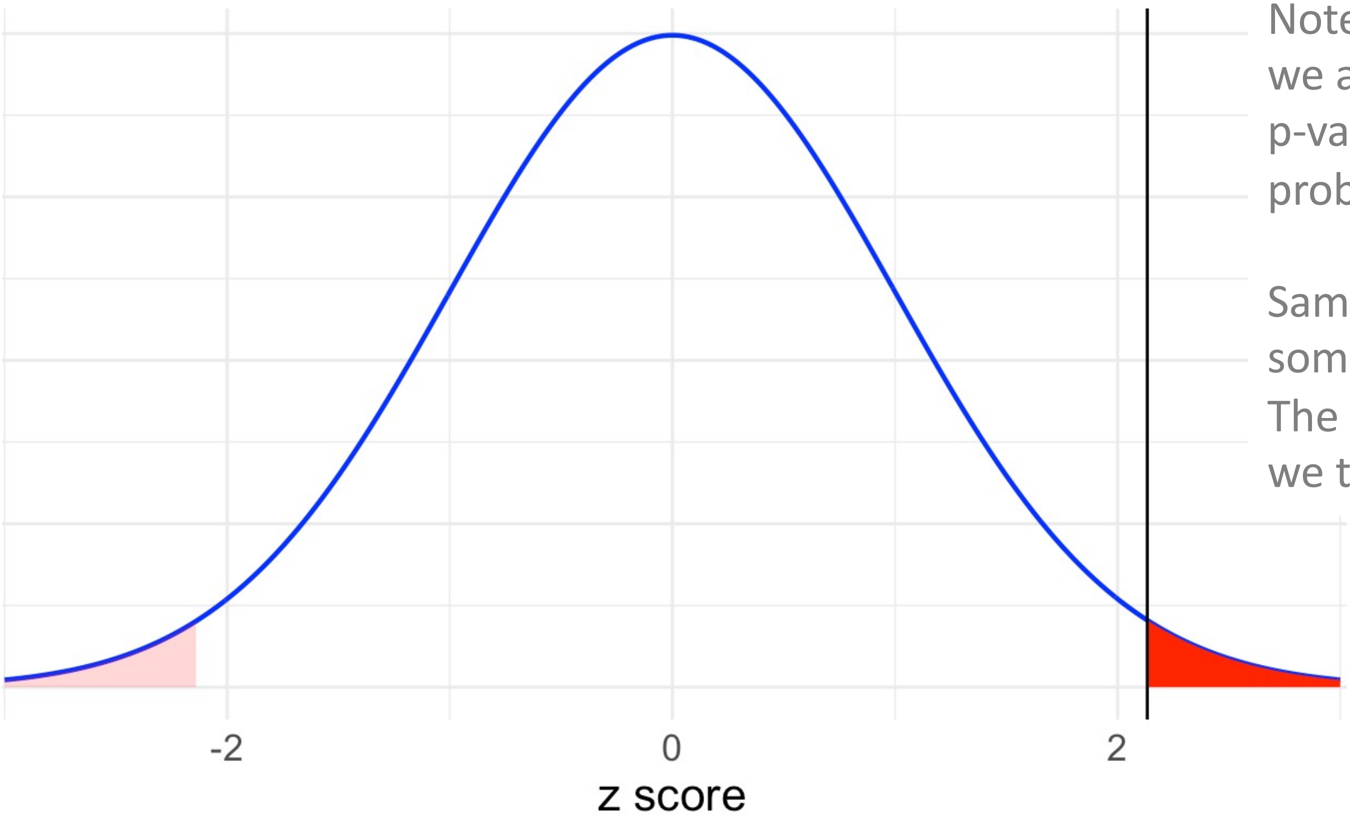
```
z = (mean(x) - 100)/15*sqrt(16)                                 [1] 2.133
Z_crit = qnorm(0.05/2)                                          [1] -1.96
```



Standard normal (Z) dist. for NHST

```
z = (mean(x) - 100)/15*sqrt(length(x))                [1] 2.133
```

```
(cur.p.val = 2*(1-pnorm(z))                            [1] 0.0329
```



Notes:
we are calculating a ‘two-tailed’ p-value by multiplying the probability of one tail by two.

Sampling variability means some slight imprecision here. The more sampled h0 means we take, the less imprecision.

NHST Z tests

- Calculate z-score of sample mean relative to null hypothesis sampling distribution of the sample mean

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{X}}^{0}}{\sigma_{\bar{X}}^{0}} = \frac{\bar{x} - \mu_{\bar{X}}^{0}}{\frac{\sigma_{X}^{0}}{\sqrt{n}}} = \left(\frac{\bar{x} - \mu_{\bar{X}}^{0}}{\sigma_{X}^{0}}\right)\sqrt{n}$$

- Classic approach: z score past significance threshold?
  - Compare z score to critical z score for alpha level.
  - Reject or retain (fail to reject) null hypothesis
- Modern approach (p-value below alpha?)
  - Calculate p-value: probability of a z-score at least as extreme as this one under the null hypothesis.
  - Compare to alpha value.

Two approaches to NHST

```
z = (mean(x) - 100)/15*sqrt(16)                    [1] 2.133
```

Classic approach: compare test statistic to critical statistic value?

```
z.crit = abs(qnorm(0.05/2))                        [1] 1.96
```

```
abs(z.score) > abs(z.crit)                         [1] FALSE
```

Modern approach: compare p-value of test statistic to alpha?

```
p.val=2*(pnorm(-abs(z.score)))                     [1] 0.033
```

```
p.val < 0.05                                       [1] FALSE
```

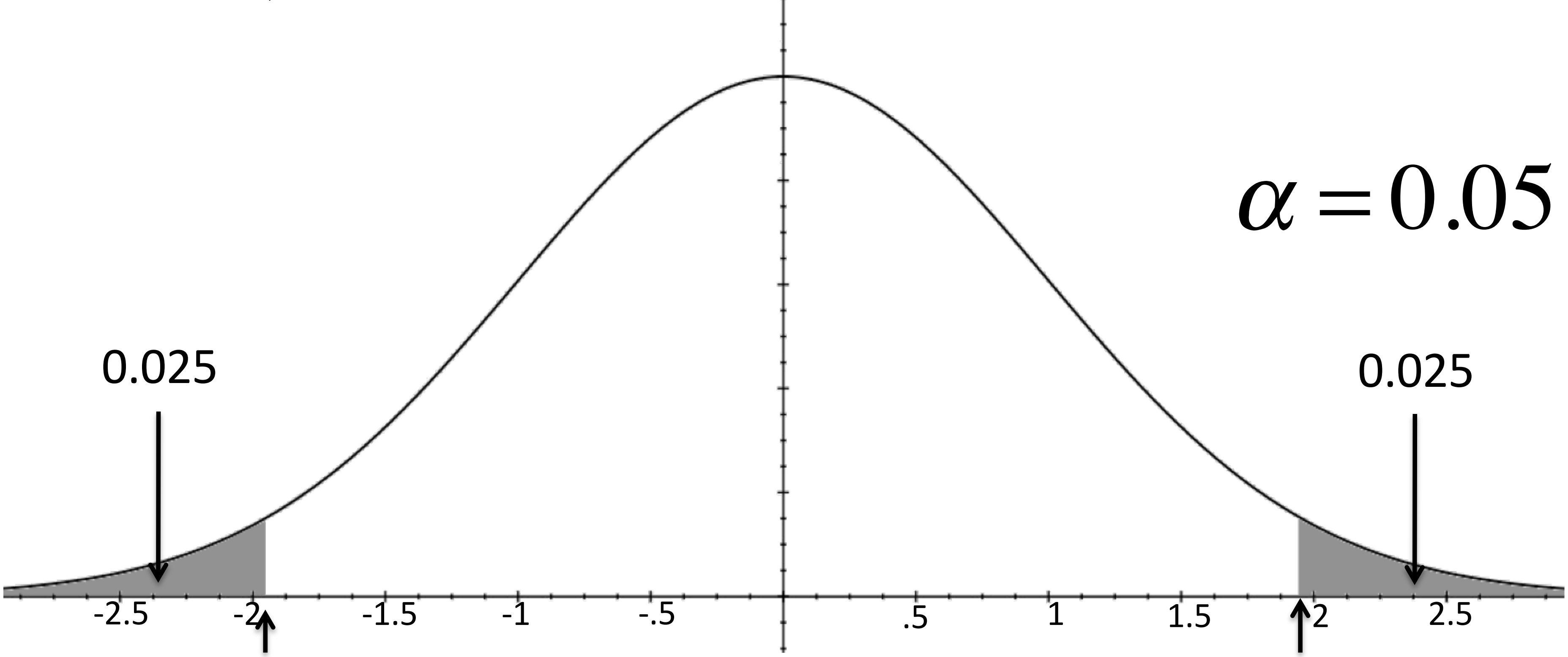
P-value is more informative, but a confidence interval is better yet

Back in the day: Probability tables

- Very hard to evaluate p-value for arbitrary Z scores.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt \qquad \Phi(x) = 0.5 + \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \left[x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \cdots + \frac{x^{2n+1}}{(2n+1)!!} + \cdots \right]$$

- Instead, find the ‘critical’ z-value



$\alpha = 0.05$

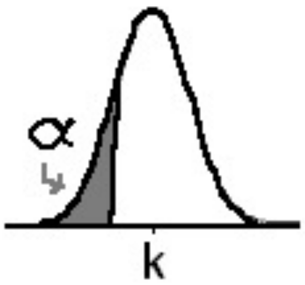
0.025

0.025

$z_{critical}$ = -1.96

$z_{critical}$ = +1.96

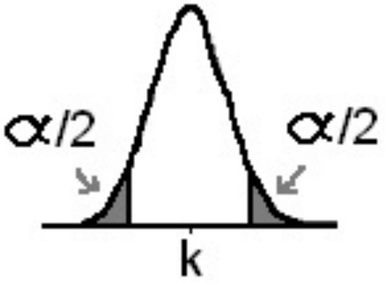
Back in the day: Probability tables.



$H_0: \mu = k$

$H_1: \mu < k$

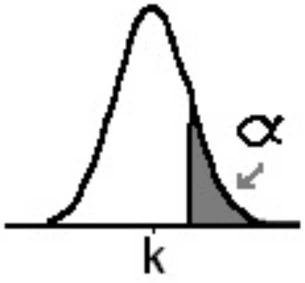
α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33



$H_0: \mu = k$

$H_1: \mu \neq k$

α	z critical
0.10	± 1.65
0.05	± 1.96
0.01	± 2.58

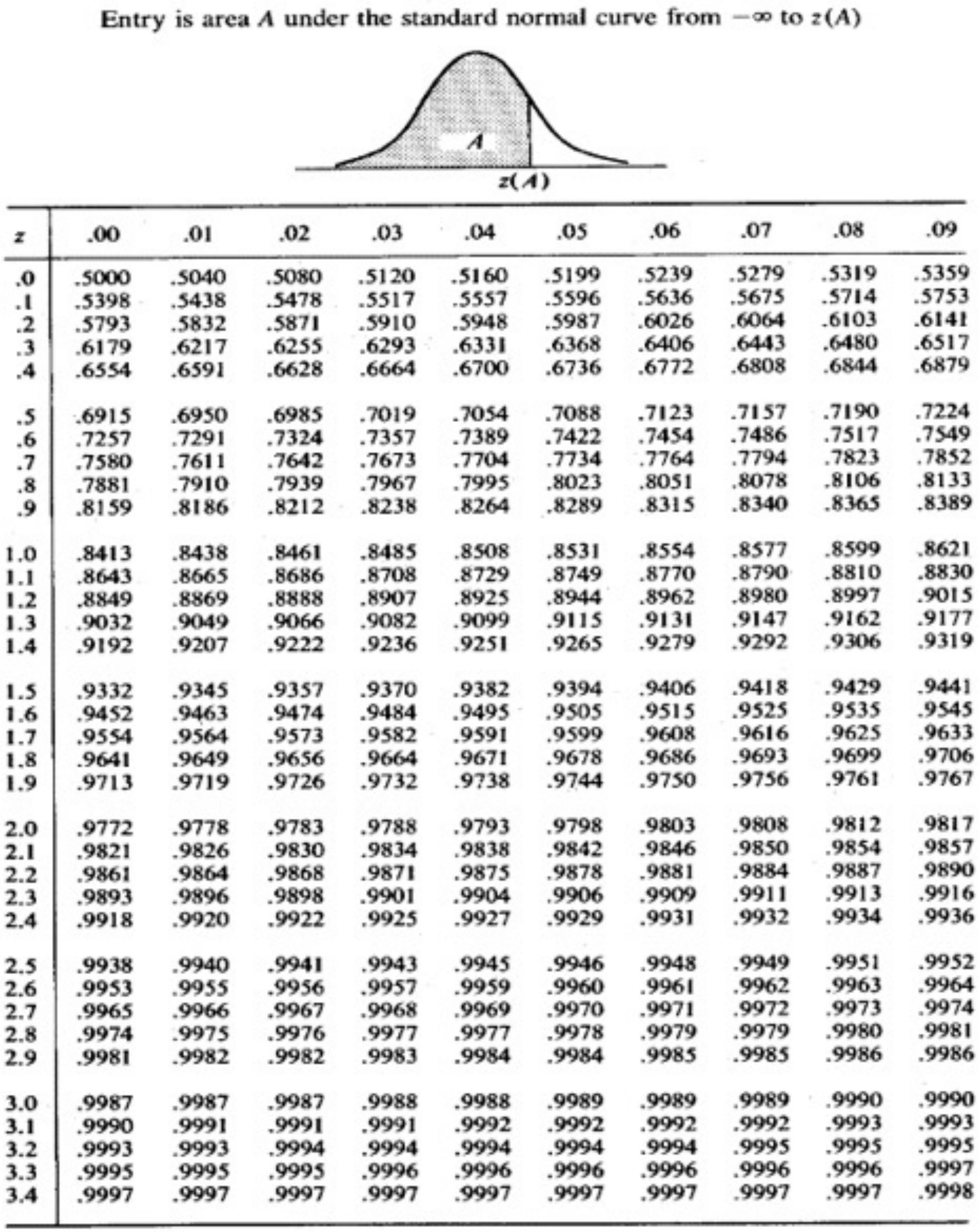


$H_0: \mu = k$

$H_1: \mu > k$

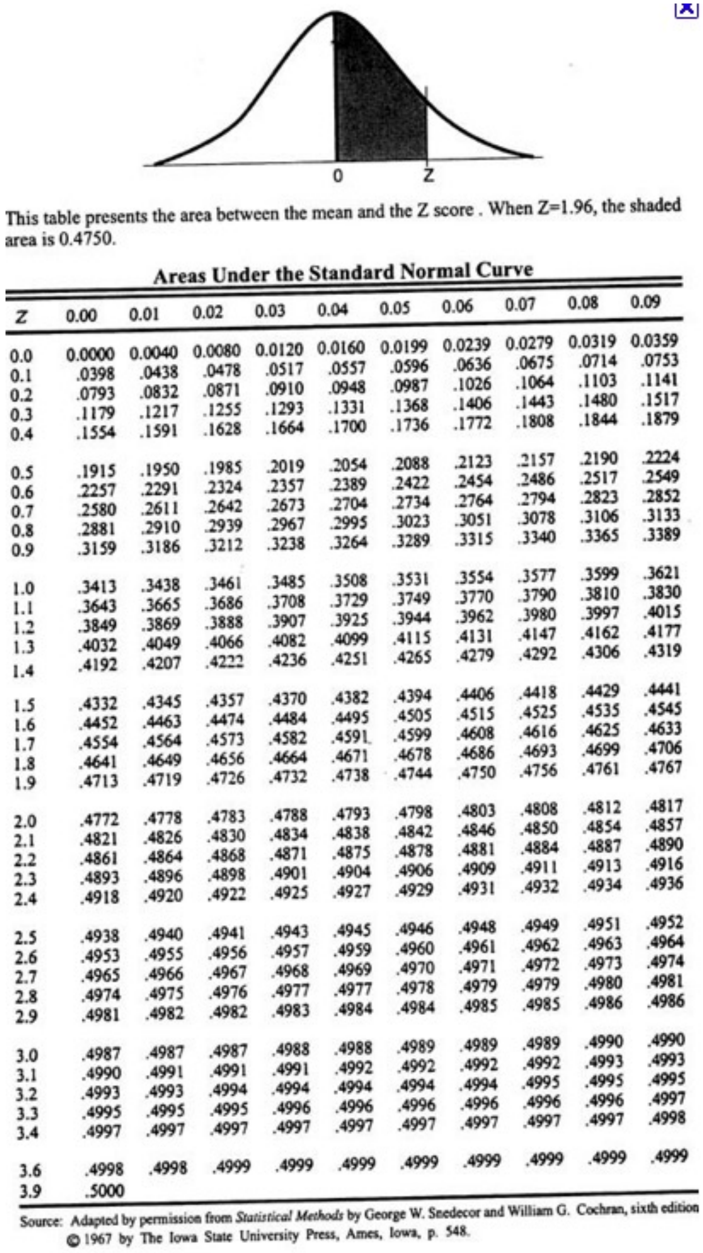
α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

Back in the day: Probability tables.



Entry is area *A* under the standard normal curve from −∞ to *z*(*A*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



This table presents the area between the mean and the Z score . When Z=1.96, the shaded area is 0.4750.

Areas Under the Standard Normal Curve

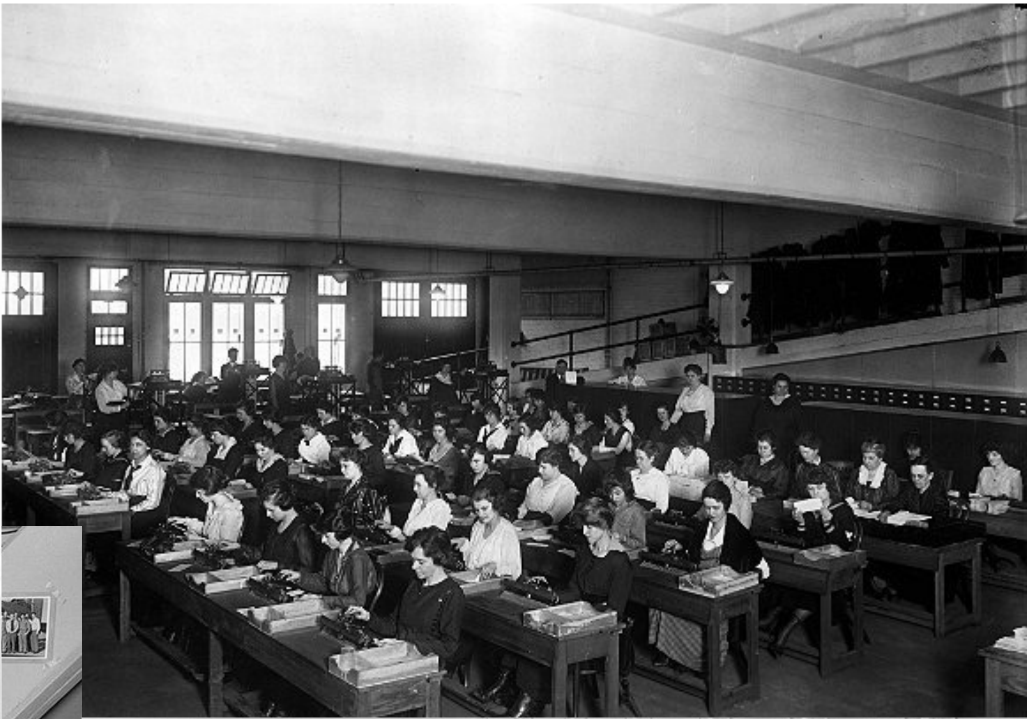
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000									

Source: Adapted by permission from *Statistical Methods* by George W. Snedecor and William G. Cochran, sixth edition © 1967 by The Iowa State University Press, Ames, Iowa, p. 548.

Where did the tables come from?

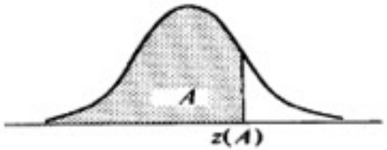
Lady computers.

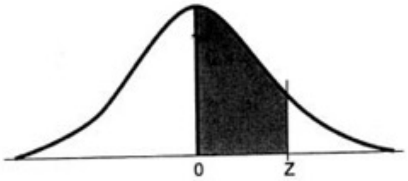
© Crown Copyright, GCHQ



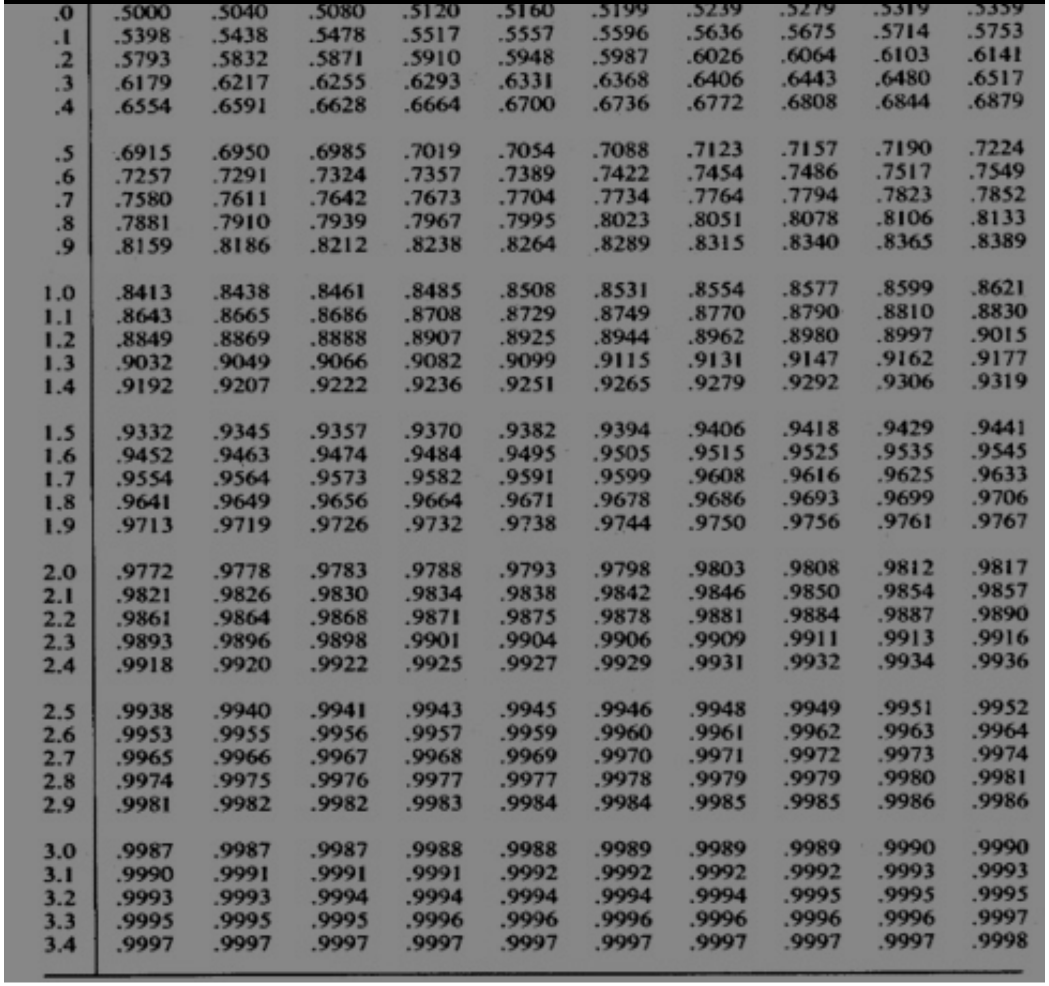
Now: pnorm

Entry is area *A* under the standard normal curve from −∞ to *z*(*A*)



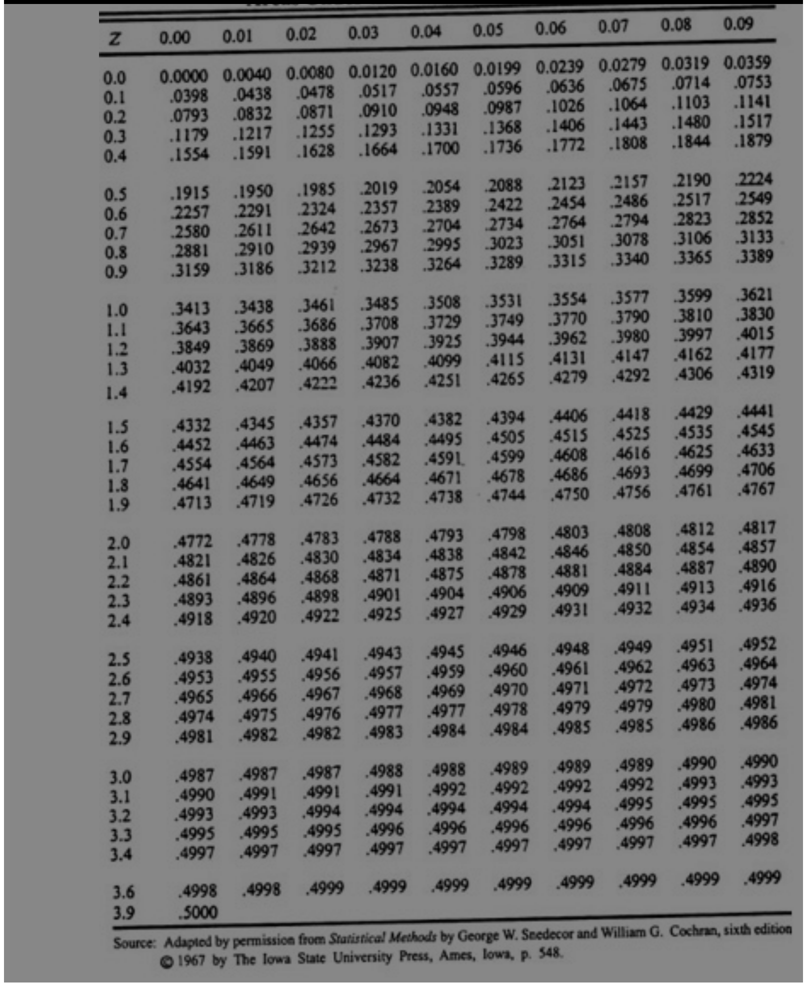


pnorm(z,0,1)



.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

pnorm(z,0,1)- 0.5

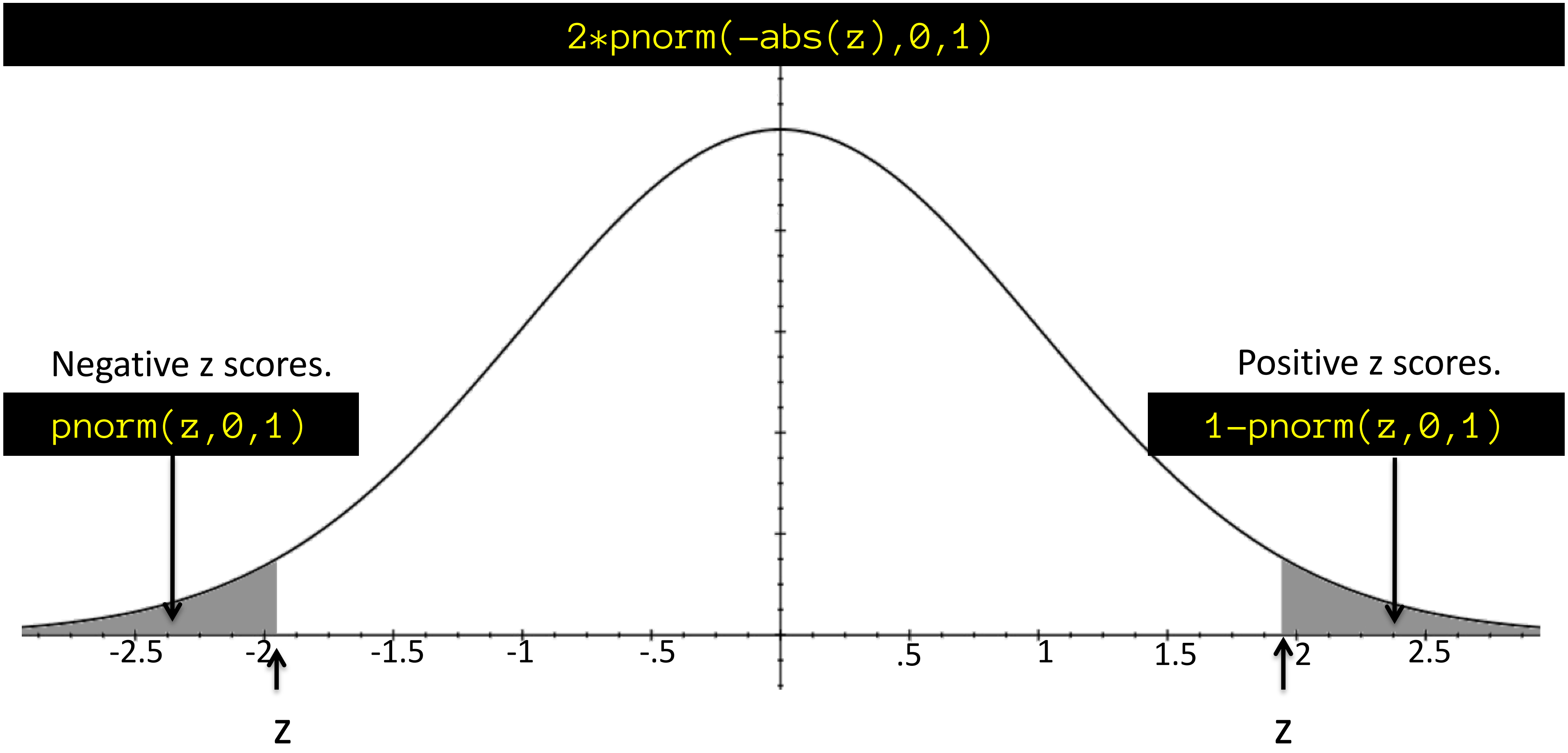


Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000									

Source: Adapted by permission from *Statistical Methods* by George W. Snedecor and William G. Cochran, sixth edition © 1967 by The Iowa State University Press, Ames, Iowa, p. 548.

Now: pnorm

- Obtain exact p-values for your actual z-score!



```
2*pnorm(-abs(z),0,1)
```

Negative z scores.

```
pnorm(z,0,1)
```

Positive z scores.

```
1-pnorm(z,0,1)
```

NHST Z tests

- Calculate z-score of sample mean relative to null hypothesis sampling distribution of the sample mean

$$z_{\bar{x}} = \frac{\bar{x}-\mu^0_{\bar{X}}}{\sigma^0_{\bar{X}}} = \frac{\bar{x}-\mu^0_{\bar{X}}}{\frac{\sigma^0_X}{\sqrt{n}}} = \left(\frac{\bar{x}-\mu^0_{\bar{X}}}{\sigma^0_X}\right)\sqrt{n}$$

- Classic approach: z score past significance threshold?
  - Compare z score to critical z score for alpha level.
  - Reject or retain (fail to reject) null hypothesis
- Modern approach (p-value below alpha?)
  - Calculate p-value: probability of a z-score at least as extreme as this one under the null hypothesis.
  - Compare to alpha value.

```
p.val = 2*pnorm(-abs(z.score))
```

Sample 16 male heights. mean=64”.

Ho: Sample from South Korean male population:

mean=68.5”   sd=4”

Can we reject Ho?

Z-score of x.bar

```
(64-68.5)/4*sqrt(16)                 [1] -4.5
```

(2-tail) p-value

```
2*pnorm(-4.5,0,1)                    [1] 6.8e-06
```

Yes, we can reject, given that our alpha level is not below ~10^{-5}

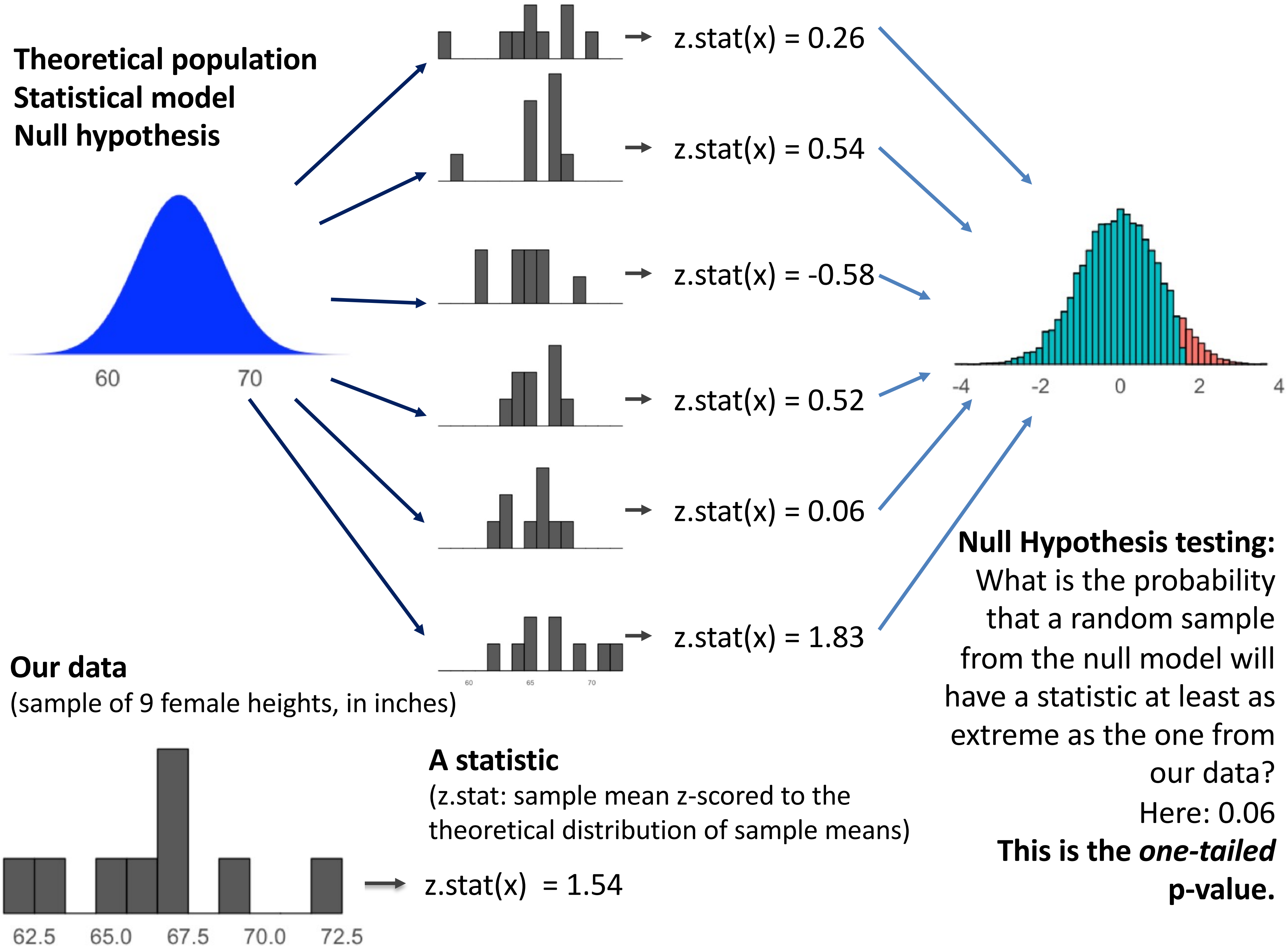
So… this sample would be very surprising if it came from the South Korean male population.

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals and NHST analogue
- Null hypothesis significance testing
- Power

Null hypothesis significance testing

- What structure are you testing?
  - Mean of sample population differs from known.
- Define a ‘null’ hypothesis lacking structure.
  - Sample drawn from known distribution.
- Define a ‘test statistic’ measuring structure
  - Z score of sample mean relative to null mean, sem
- Obtain null test stat. ‘sampling distribution’
  - Standard normal distribution.
- Compare sample statistic to H_o distribution
  - Obtain p-value, compare to α
  - Reject or fail to reject the null hypothesis.



Theoretical population
Statistical model
Null hypothesis

z.stat(x) = 0.26

z.stat(x) = 0.54

z.stat(x) = -0.58

z.stat(x) = 0.52

z.stat(x) = 0.06

z.stat(x) = 1.83

Our data
(sample of 9 female heights, in inches)

A statistic
(z.stat: sample mean z-scored to the theoretical distribution of sample means)

z.stat(x) = 1.54

Null Hypothesis testing:
What is the probability that a random sample from the null model will have a statistic at least as extreme as the one from our data?
Here: 0.06
This is the *one-tailed* p-value.

Some jargon.

- **Null hypothesis (H_o)**
  a model of the data that lacks structure we want to test for – the "boring" alternative.
- **Test statistic**
  a sample statistic that measures the structure we wish to test for.
- **Sampling distribution of test statistic under H_o**
  the probability distribution over test statistics we expect to see in samples from the null hypothesis.
- **P-value**
  the probability of observing a test-statistic at least as extreme as ours under the H_o distribution. (~surprise)
- **α-level**
  How often are we willing to falsely reject the null hypothesis?
  (our chosen prob. of Type I error)

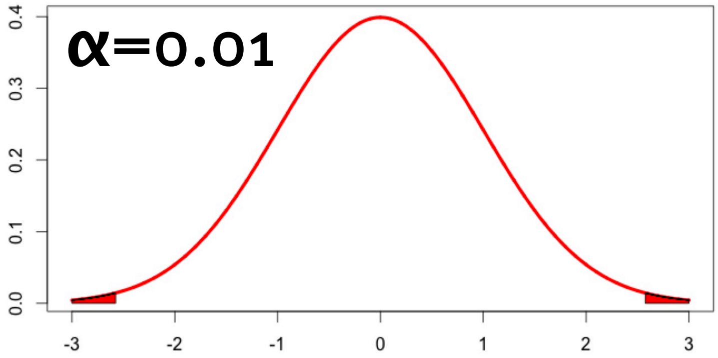
NHST terms & concepts

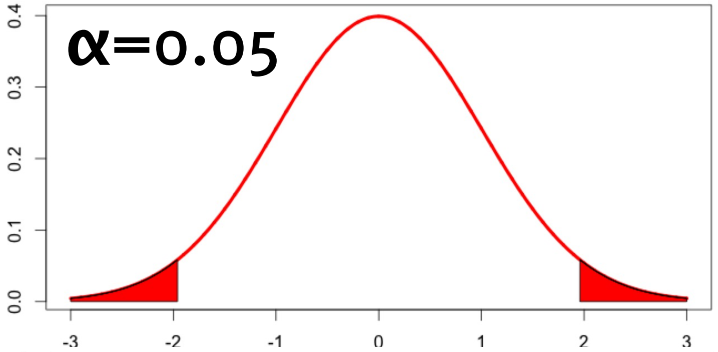
- Null hypothesis
- ‘Alternative hypothesis’
- Type I error
- Type II error
- Alpha
- P-value
- Effect size
- Power

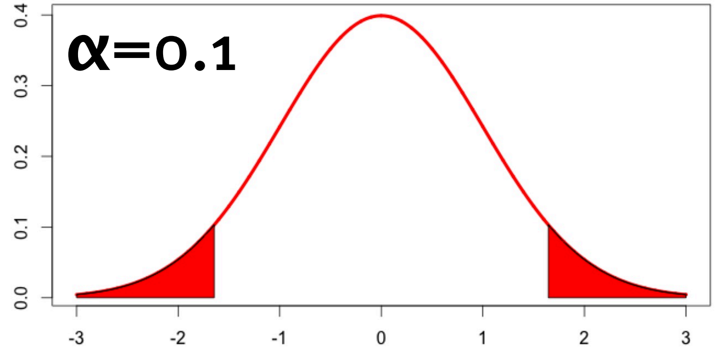
H_0 (null hypothesis) testing errors

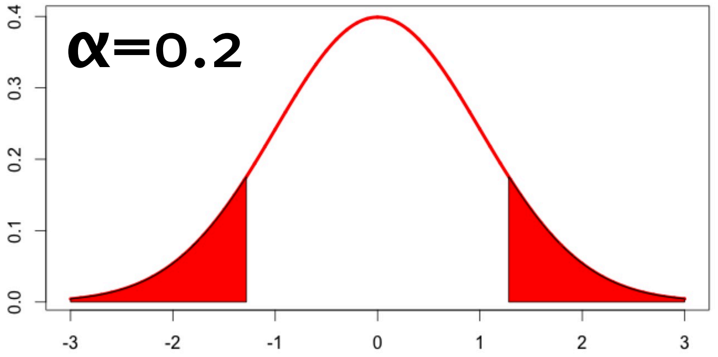
		Reality	
		H_0 false	**H_0 true**
Decision	Reject H_0	Woohoo!	**Type I error (false alarm)**
	Fail to reject H_0	**Type II error (miss)**	Ok.

Frankly: I always had a hard time remembering which one is Type I and which one is Type II. More important to remember the conceptual difference than the arbitrary naming convention.









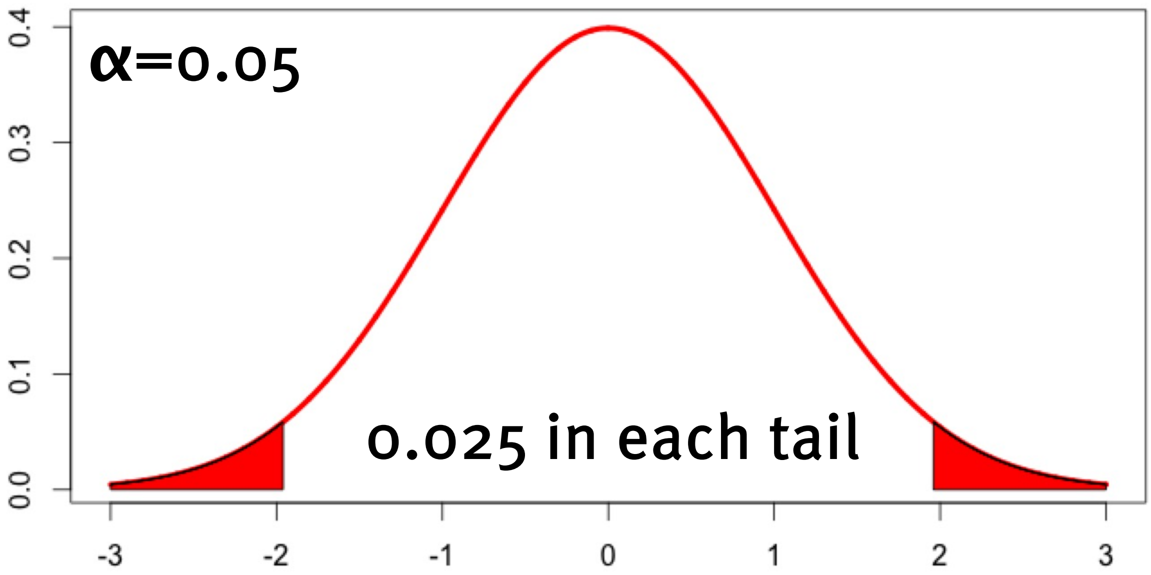
Alpha (α)

- The probability of “Type I error”: The conditional probability of falsely rejecting the null when it is actually true, that we are willing to tolerate.
- Typically we use a “two-tailed” test, which means that this probability is equally distributed into the two tails.
- This determines which values we are willing to reject the null for.

Alpha (α)

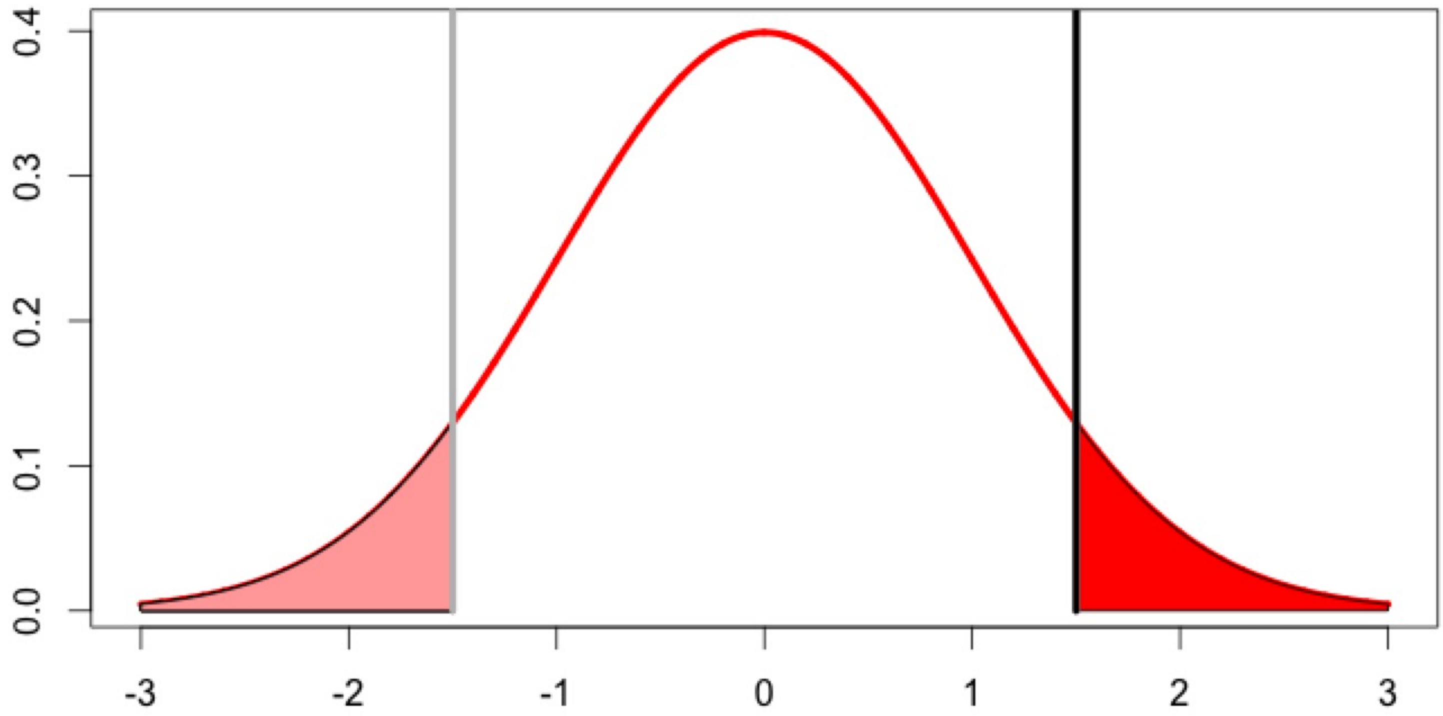
- In theory, we decide on alpha. In practice, we just follow the 5% (0.05, 1/20) convention for better or ***worse**:**

  **I endorse recent push to lower convention. Be wary of $0.01 < p < 0.05$.*

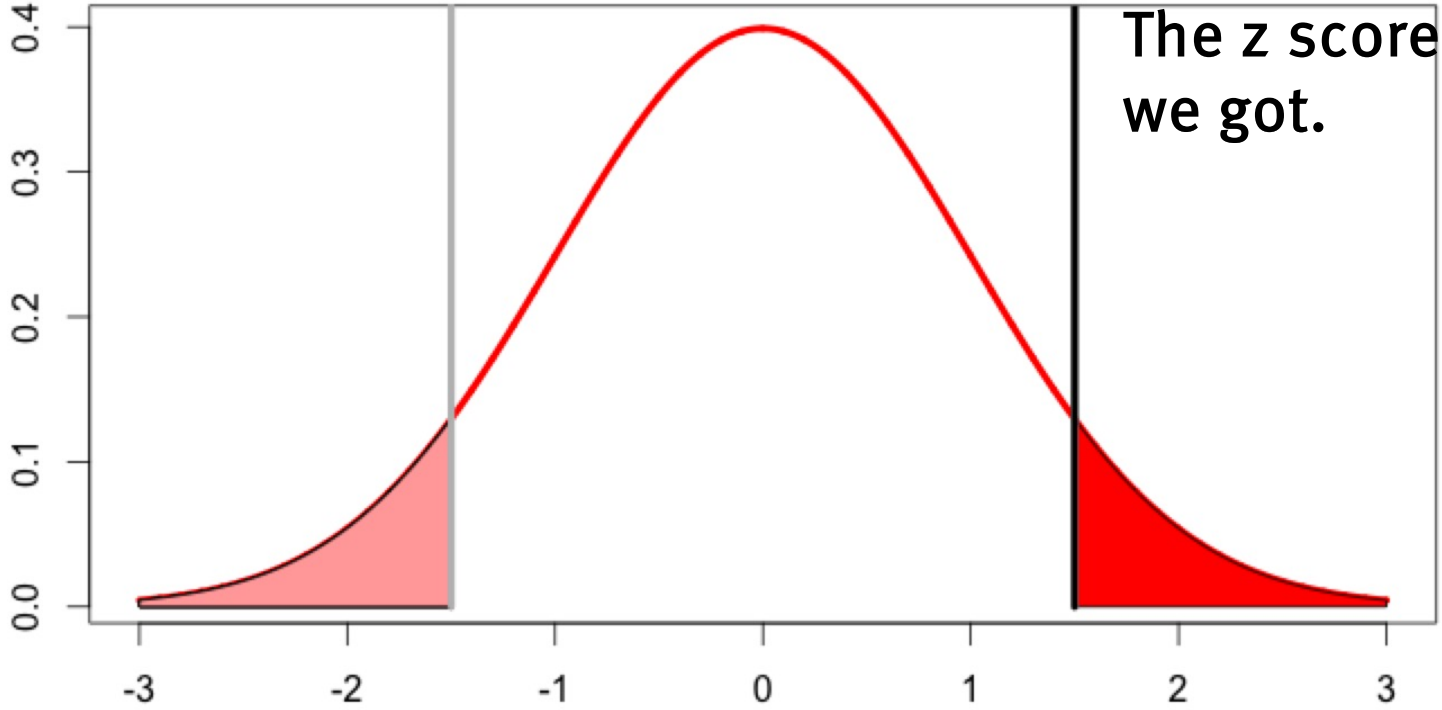


P-value

- The lowest alpha value at which our data would reject the null hypothesis.
- The probability of seeing an outcome at least as extreme as the one that we saw when sampling from null.
- Typically this is two-tailed.



P-value



The z score we got.

This is the second tail: the probability of seeing a z-score at least this extreme in the *other* direction.

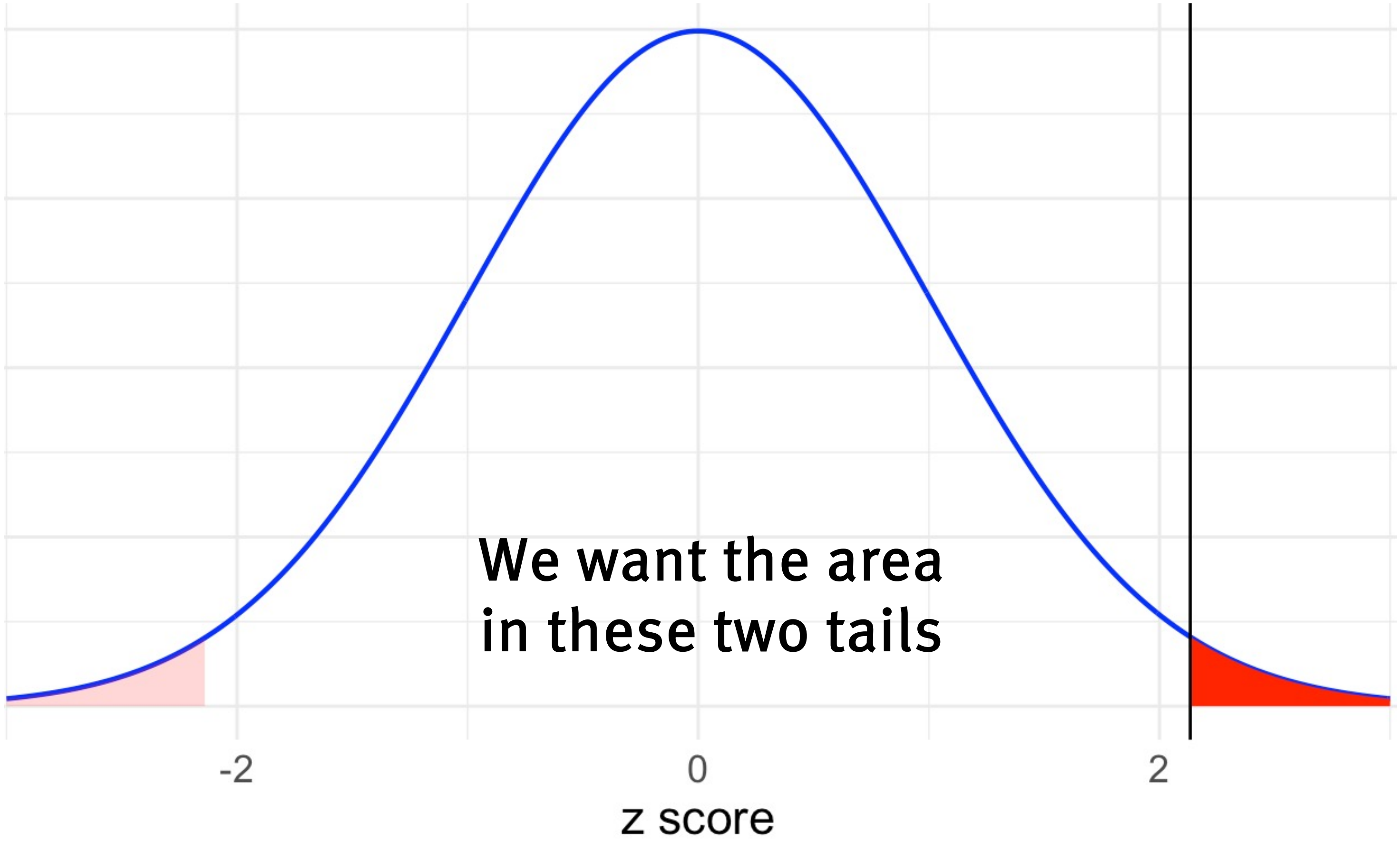
This area under the curve corresponds to the probability of seeing a z-score at least this extreme in *this* direction.

These two tails together give us the p-value.
(due to symmetry, we can just multiply one tail by 2)

P-value

```
z = (mean(x) - 100)/15*sqrt(16)                    [1] 2.133
Z_crit = qnorm(0.05/2)                             [1] -1.96
```

- What is the p-value of the null hypothesis that 201 IQs are samples from the overall IQ distribution?



We want the area in these two tails

P-value

```
z = (mean(x) - 100)/15*sqrt(16)                    [1] 2.133
Z_crit = qnorm(0.05/2)                             [1] -1.96
```

- What is the p-value of the null hypothesis that 201 IQs are samples from the overall IQ distribution?

```
P_upper = 1-pnorm(2.133)                           [1] 0.01646
```

```
P_lower = pnorm(-2.133)                            [1] 0.01646
```

```
p.value = P_upper + P_lower                        [1] 0.03292
```

```
p.value = 2*pnorm(-abs(z))                         [1] 0.03293
```

P-value

```
z = (mean(x) - 100)/15*sqrt(16)                    [1] 2.133
Z_crit = qnorm(0.05/2)                             [1] -1.96
```

- What is the p-value of the null hypothesis that 201 IQs are samples from the overall IQ distribution?

```
p.value = 2*pnorm(-abs(z))                       [1] 0.03292
```

- This means:
  we could reject this null hypothesis at alpha=0.03292
  and
  the probability of seeing an outcome at least as extreme as ours under the null hypothesis is 0.03293

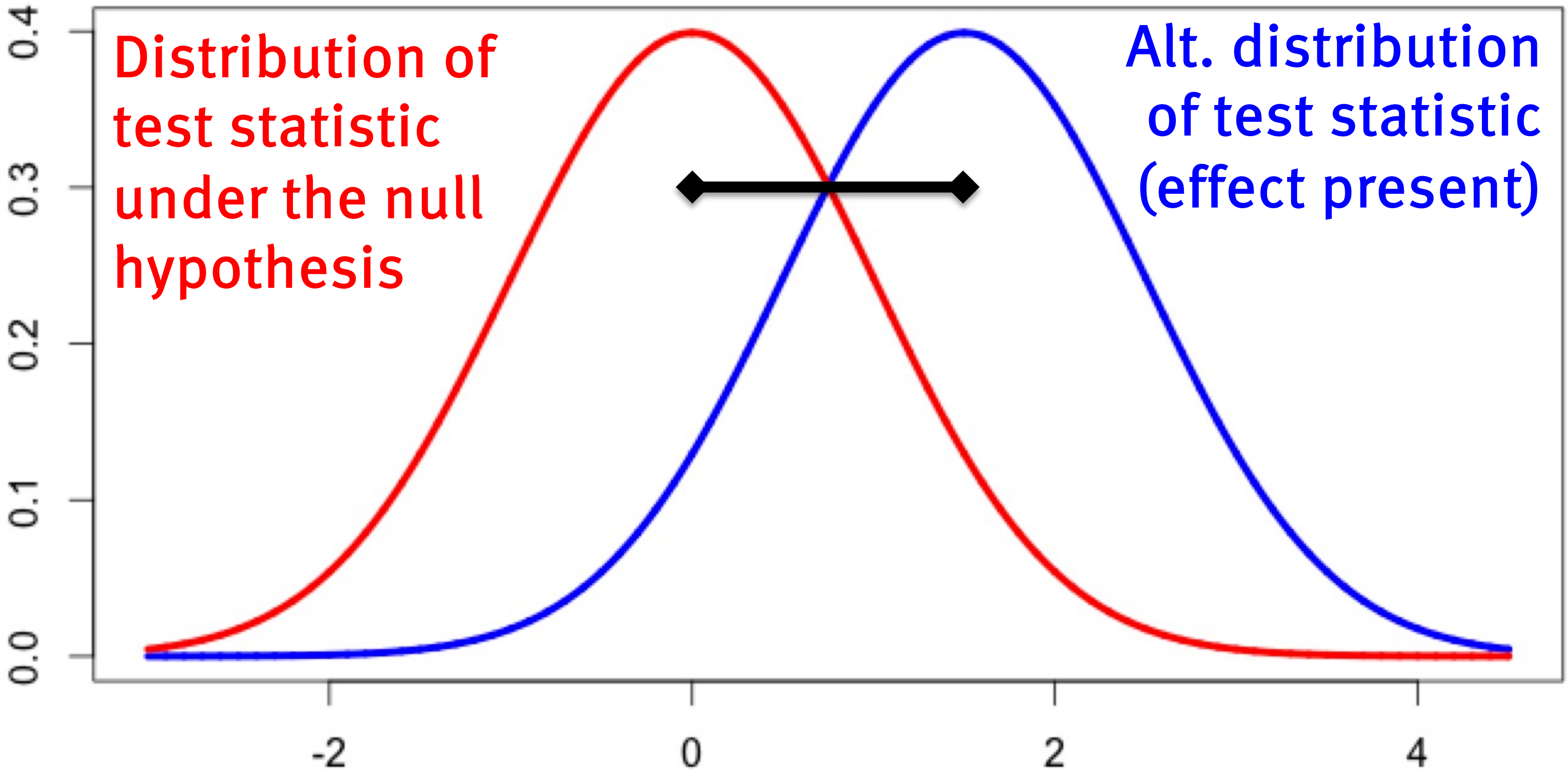
Alternative hypothesis

- This might be just “the null hypothesis is false” Meaning: whatever structure we tested for *is* there.

- E.g.: the mean IQ *is* different for 201 students compared to the overall population

- However, if you want to assess the *power* of your test (which, is very important), the alternate hypothesis needs to be specified as an actual distribution.

Alternative hypothesis



Distribution of test statistic under the null hypothesis

Alt. distribution of test statistic (effect present)

“Effect size”: difference between the alternative and null distributions.
(often: difference in means, in units of standard deviation)

Typically when considering effect size the alternative we consider is the “true” distribution

Effect size.

- How big is this difference from the null that we are measuring?
  - (True mean - Null mean)? *Hard to compare across measures*

    $$\Delta = \mu_T - \mu_0$$

  - Z score based on sampling dist of sample mean? *Will depend on sample size, which we don't want.*

    $$z_{\bar{x}} = \frac{\mu_T - \mu_0}{\sigma_X}\sqrt{n}$$

  - **Cohen's d** (mean difference scaled by standard deviation)
    (usually we use absolute values)

    $$d = \left|\frac{\mu_T - \mu_0}{\sigma_X}\right|$$

- Sample mean often used to estimate effect size.

  $$\hat{d} = \left|\frac{\bar{x} - \mu_0}{\sigma_X}\right|$$

Effect size.

There are three practical uses of effect size:

(1) You *know* the properties of true population, and you want to know how big that difference is (perhaps to use to calculate future power analyses for unknown samples – the prototypical quality assurance scenario for NHSTs)

$$d = \left| \frac{\mu_T - \mu_0}{\sigma_X} \right|$$

(2) You have an idea what difference would be scientifically/practically meaningful, so you postulate a minimum relevant effect size.

(3) You ran an experiment, and want to know what to expect in a future replication. Then you use the sample mean from the pilot to *estimate* the effect size.

$$\hat{d} = \left| \frac{\overline{x} - \mu_0}{\sigma_X} \right|$$

Sample 16 male heights. mean=64”.

Ho: Sample from South Korean male population:

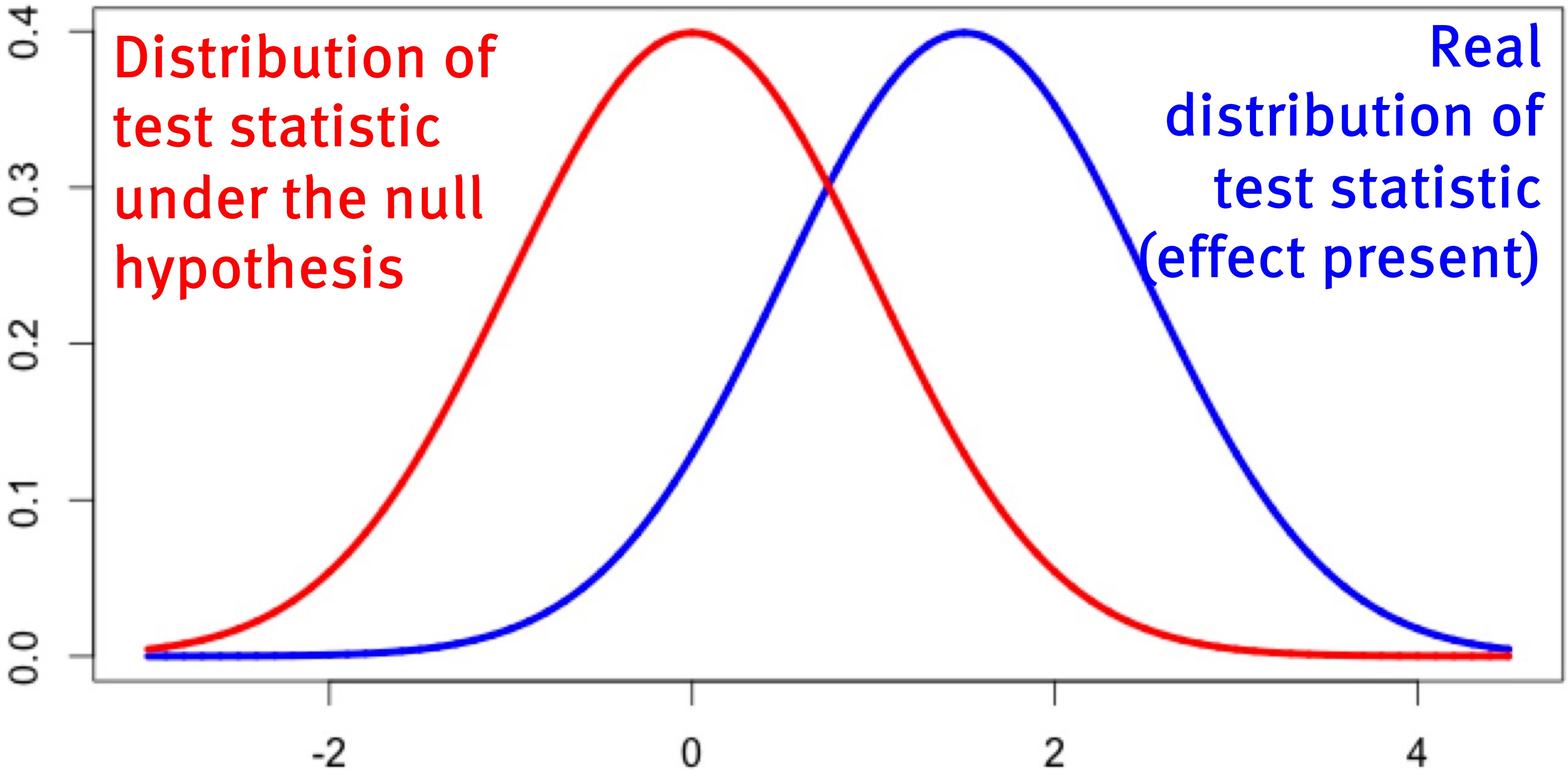
mean=68.5” sd=4”

What’s our estimate of the effect size?

$$\hat{d} = \left|\frac{\bar{x} - \mu_0}{\sigma_X}\right| = \left|\frac{64 - 68.5}{4}\right| = 1.125$$

***Estimated* effect size.**
“Estimated” because we are using the sample mean to estimate the population mean of the true effect distribution

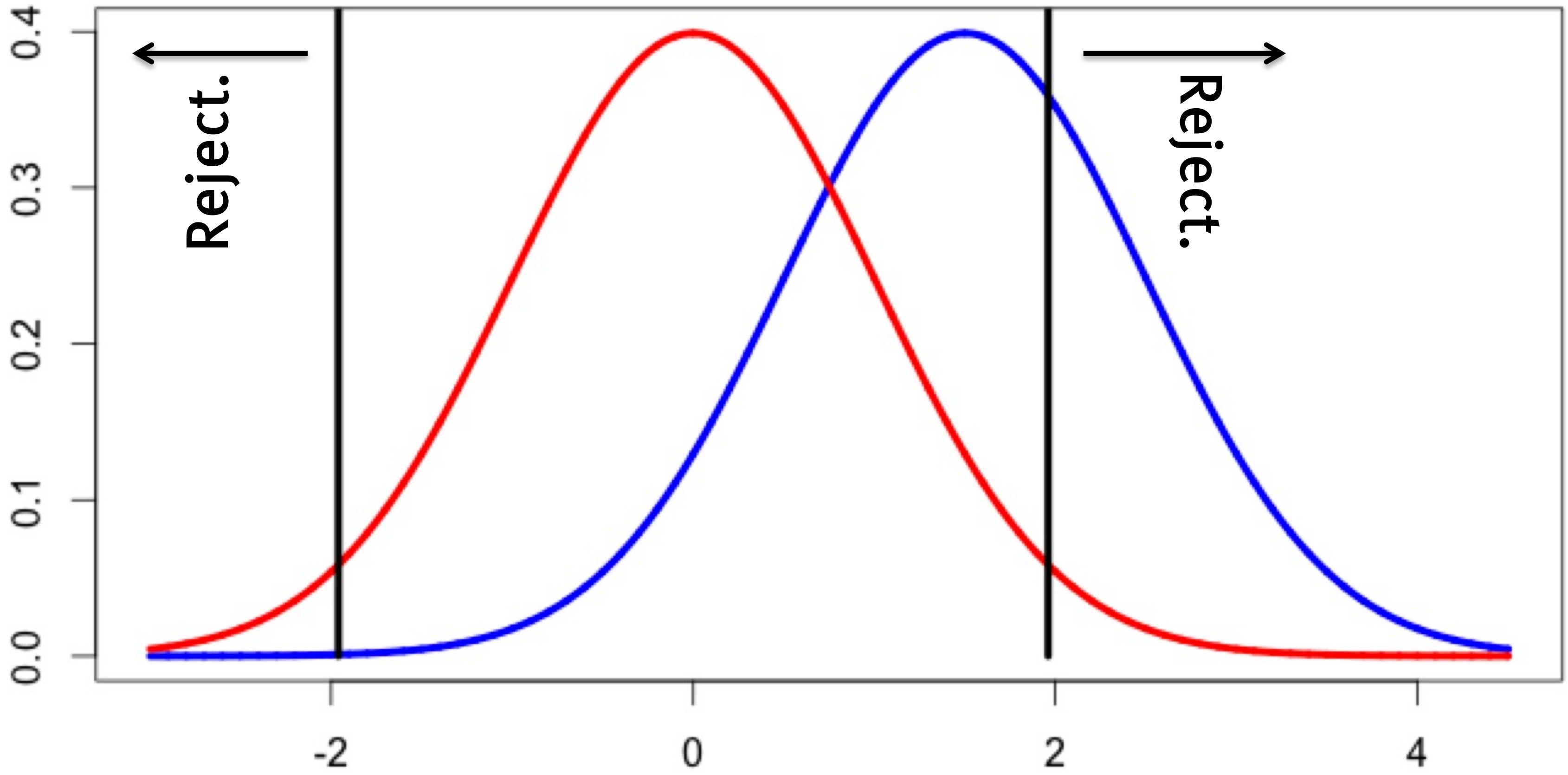
More on Type I and Type II



Distribution of test statistic under the null hypothesis

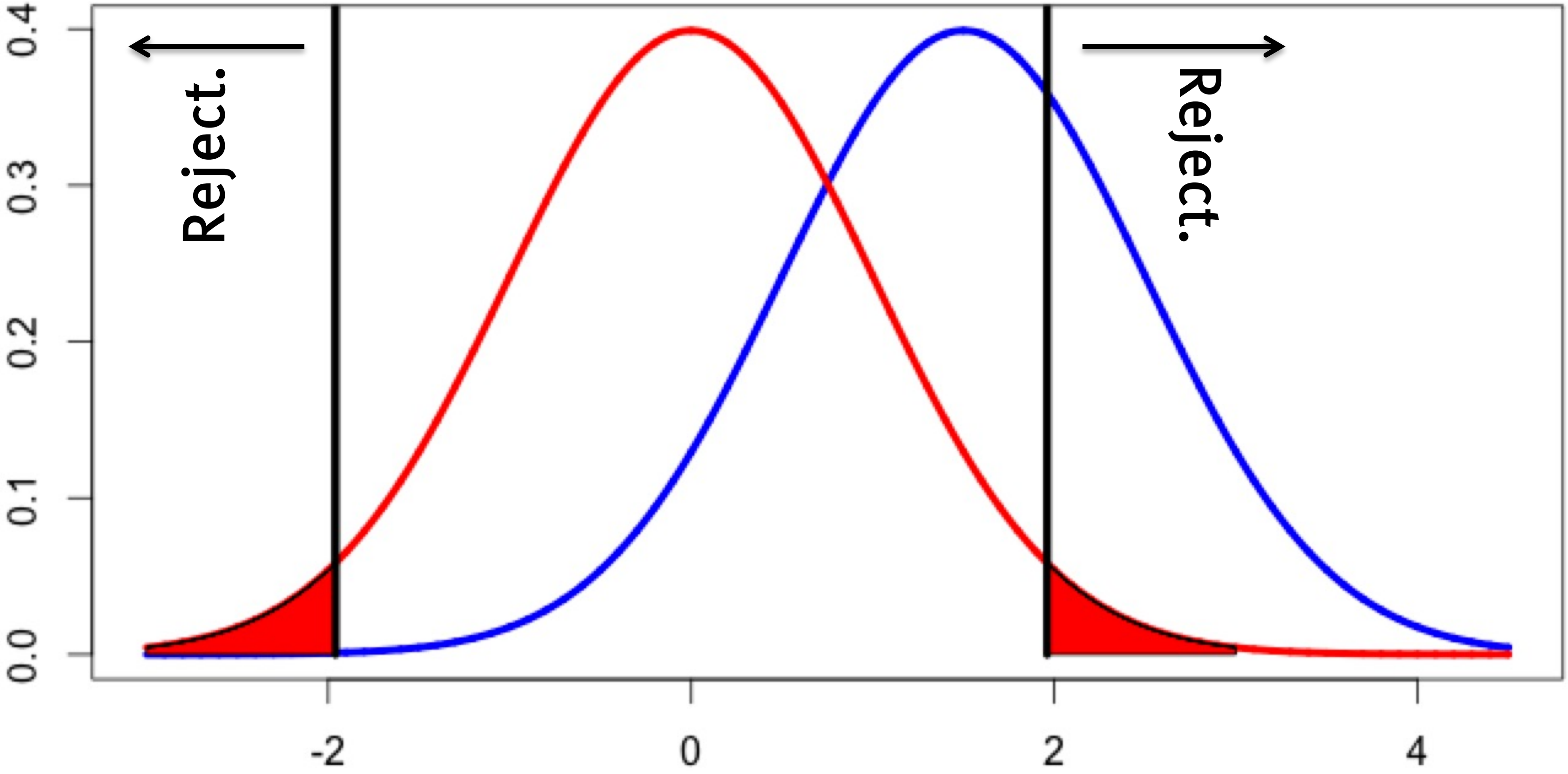
Real distribution of test statistic (effect present)

More on Type I and Type II



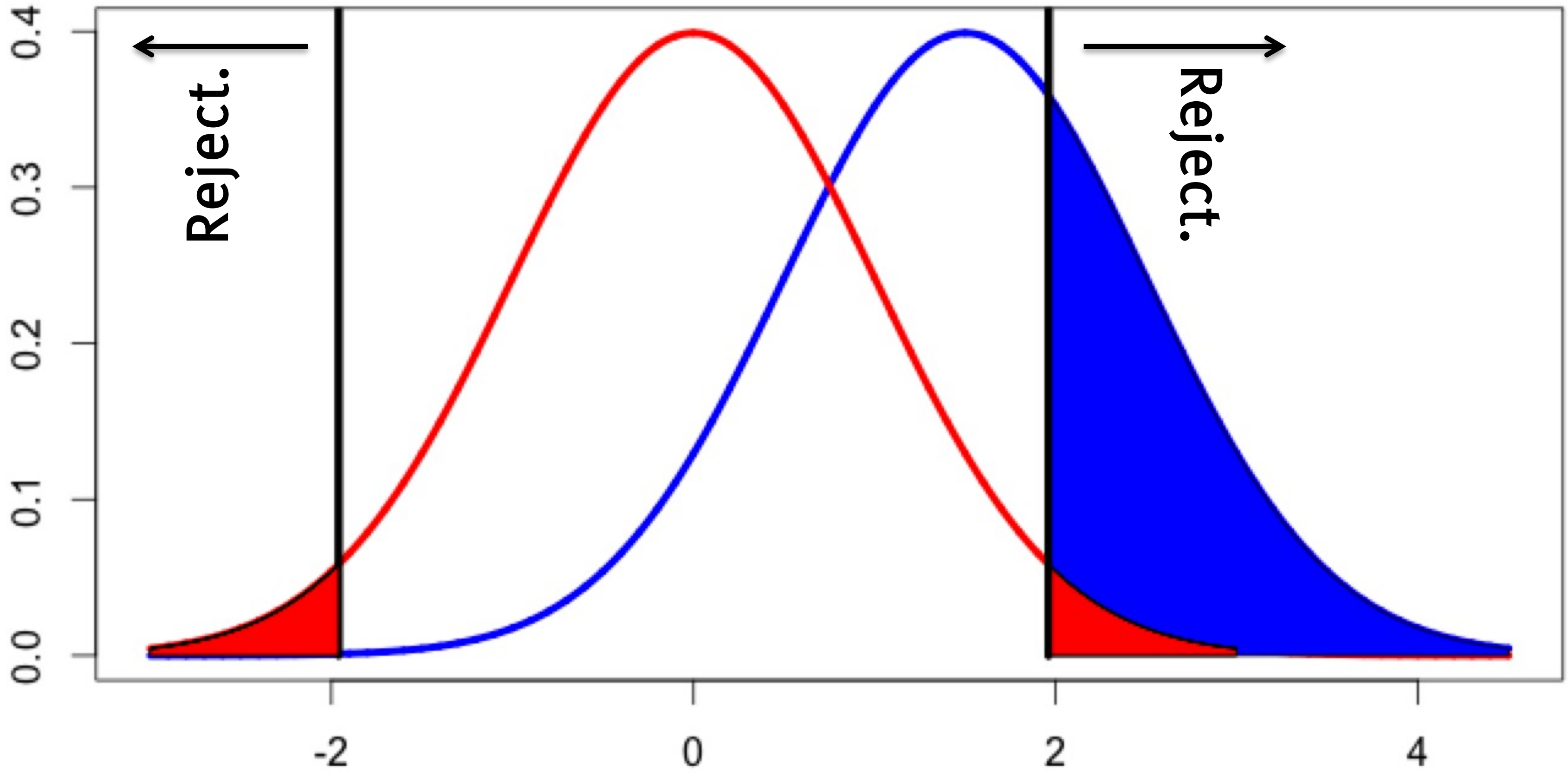
We will reject the null hypothesis if the test statistic we get from our data is past these 'critical' values
(here z_crit = -1.96 and 1.96, corresponding to alpha=0.05)

More on Type I and Type II



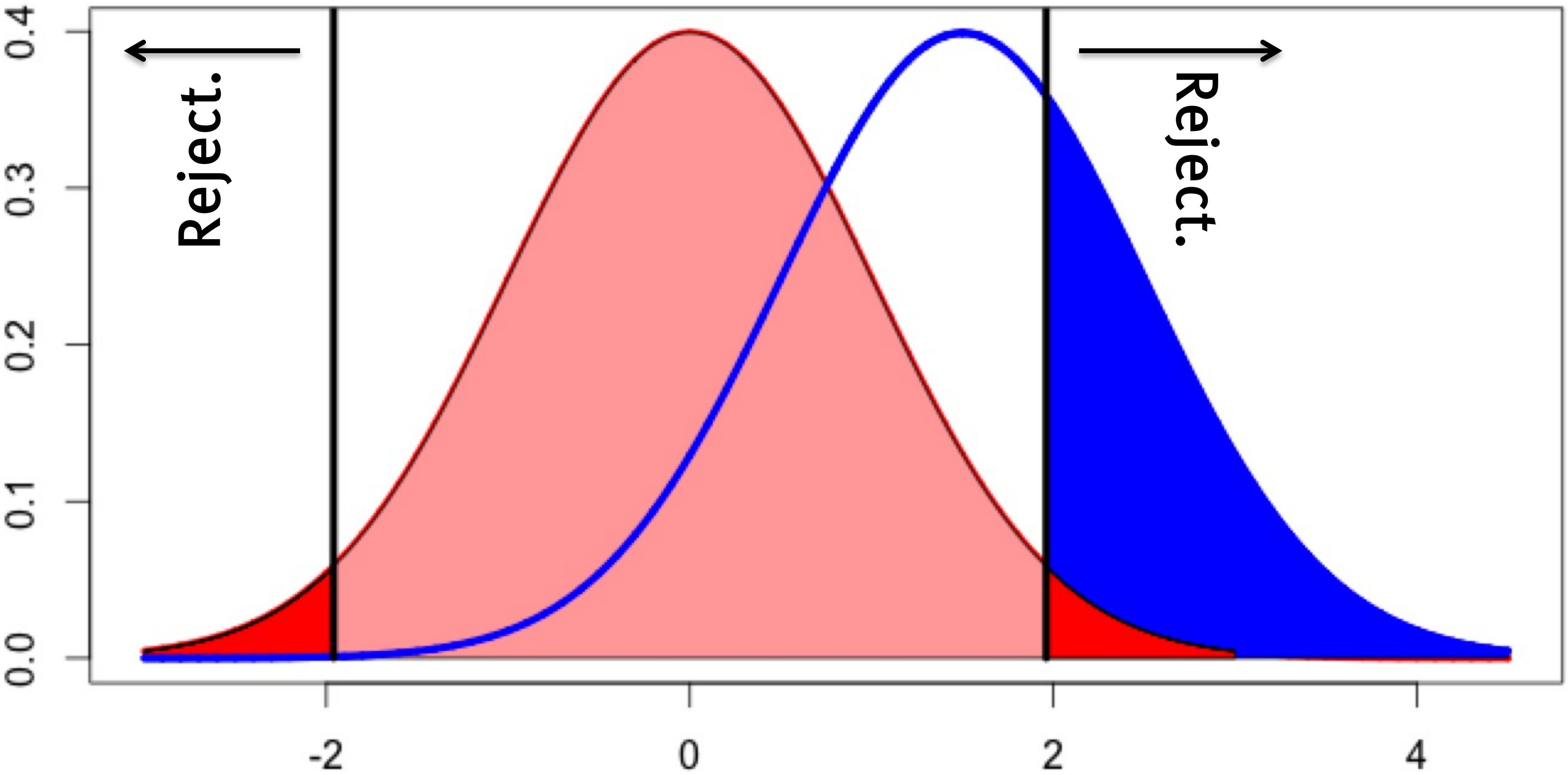
Shaded red area is the probability of Type I error:
The probability that we will reject the null hypothesis for a test statistic that actually came from the null hypothesis.
This is alpha. This is how the critical values were defined.

More on Type I and Type II



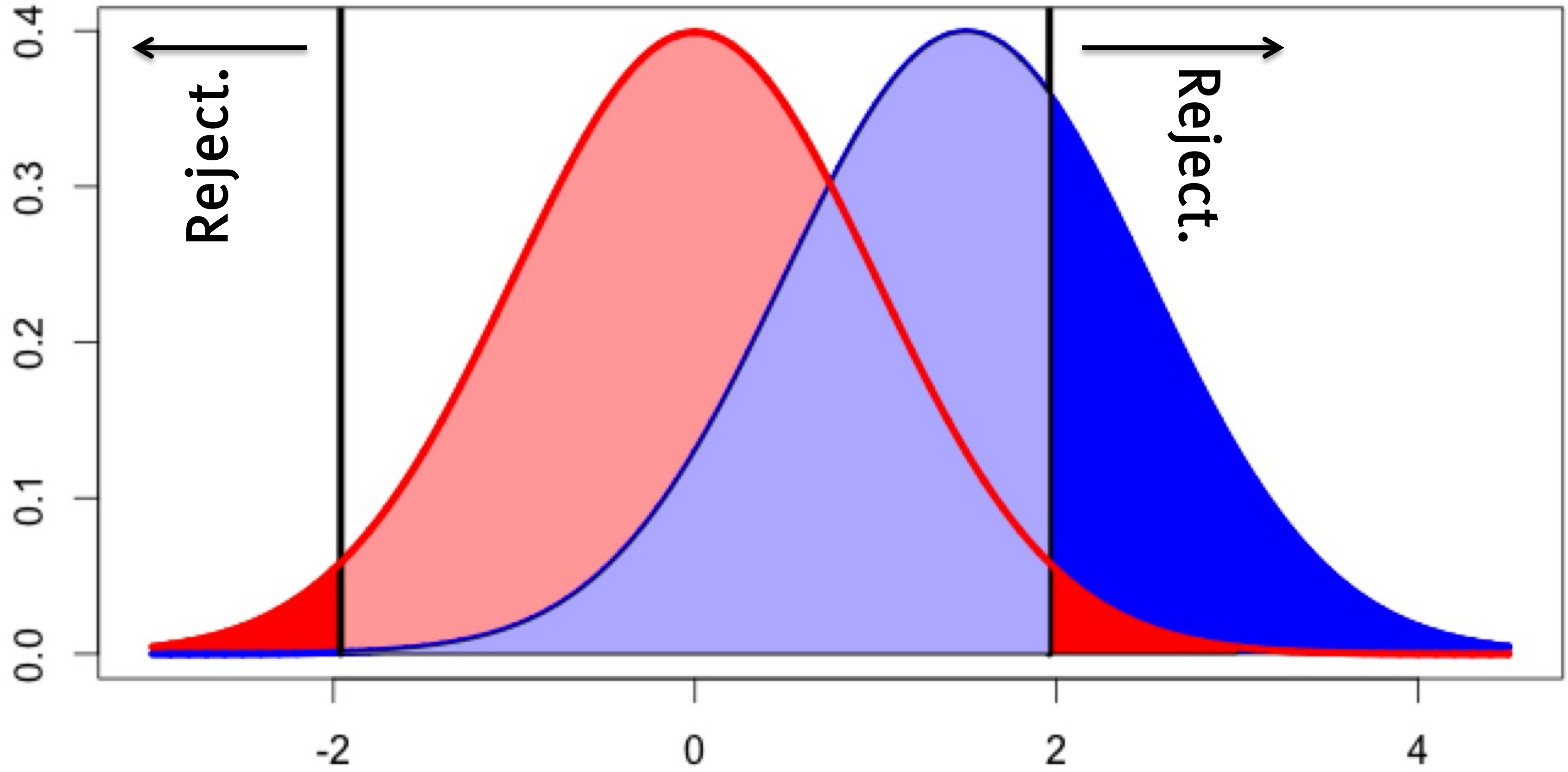
Shaded blue area is the probability of correctly rejecting the null hypothesis: rejecting the null hypothesis when the test statistic actually came from the alternative hypothesis distribution (this is "Power")

More on Type I and Type II



Shaded light red area: probability of correctly not rejecting the null hypothesis: not rejecting the null hypothesis when it is true. = 1-alpha

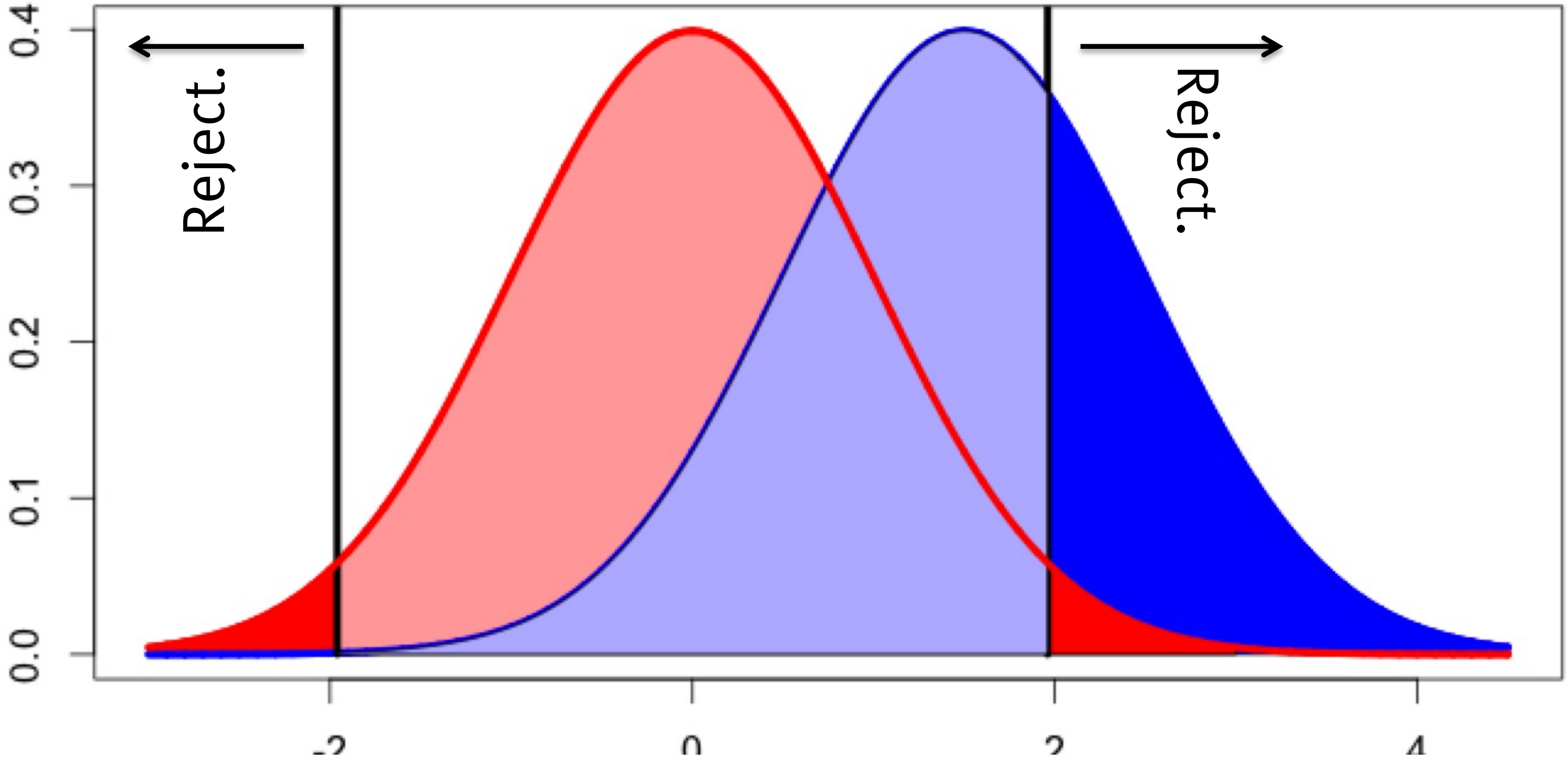
More on Type I and Type II



Shaded light blue area: probability of Type II error: Failing to reject the null hypothesis when it is actually false (when the test statistic actually came from the alternate distribution!)

This is usually called Beta

More on Type I and Type II



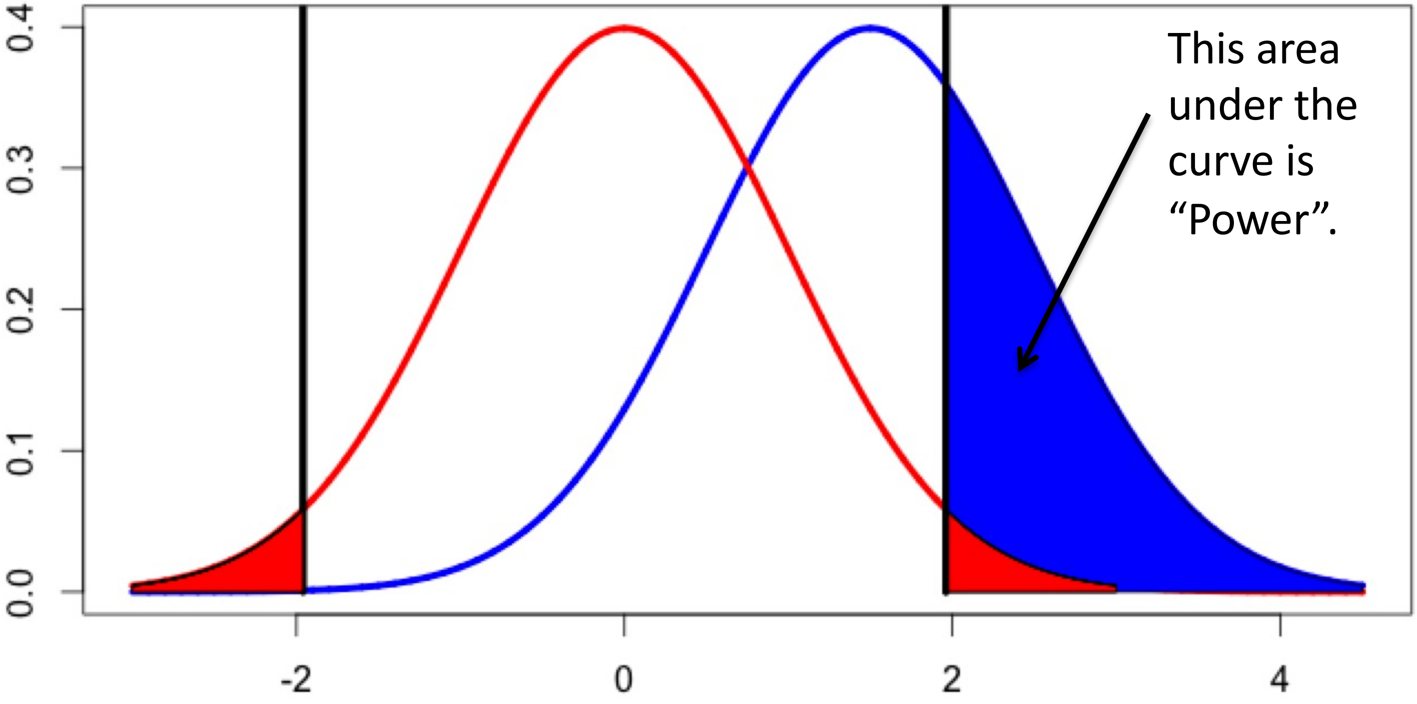
	H_0 false	H_0 true
Reject H_0	Correct rejection of null (Pr = 1-β; ‘power’)	Type I error (Pr = α)
Fail to reject H_0	Type II error (Pr = β)	Correct failure to reject null (Pr = 1-α)

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals and NHST analogue
- Null hypothesis significance testing
- Power

Power P(significant | not null)

- The conditional probability of rejecting the null hypothesis when the data actually came from the 'alternate' hypothesis distribution.
- To calculate this, we need to know what the 'true effect' distribution is. Usually, we just need the 'effect size'



This area under the curve is "Power".

How to get more power?

- Bigger difference between means.
- Less population variance.

Effect Size

- Bigger sample size
- Higher (closer to 1) alpha.
  - Directional tests

Sample 16 male heights from North Korea

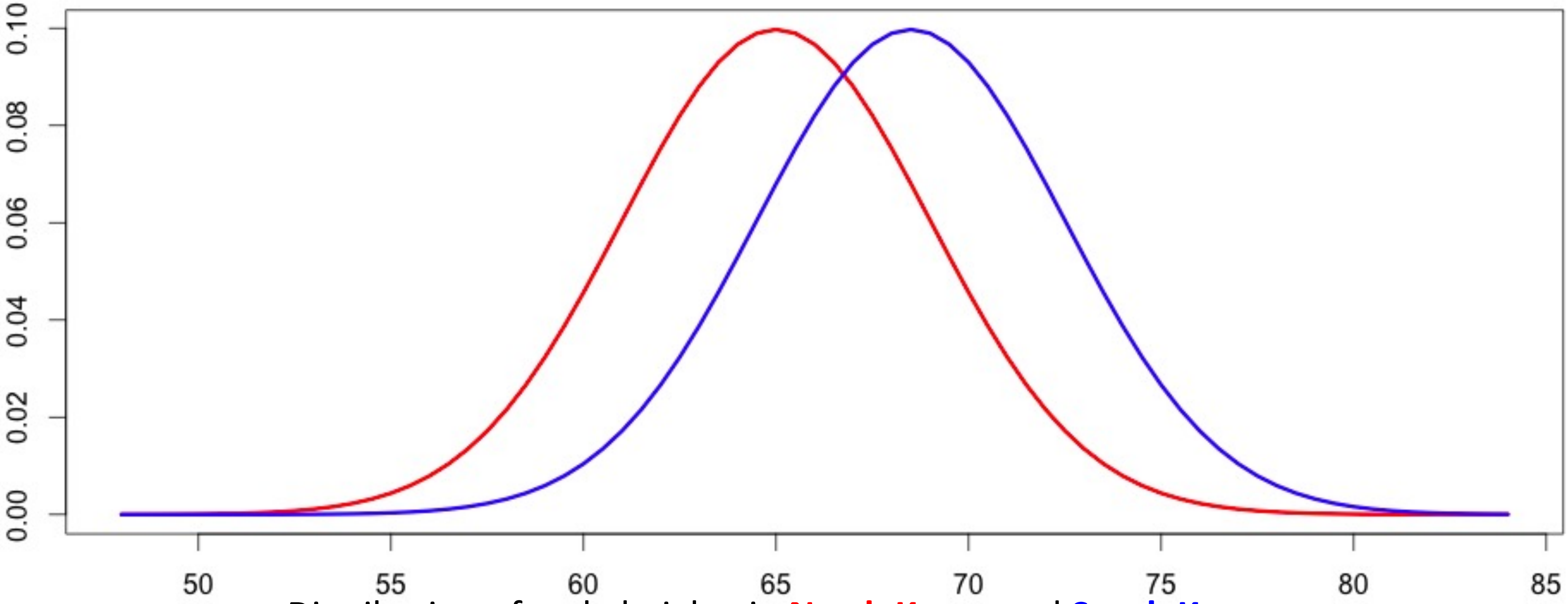
mean=65”　　sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5”　sd=4”

What is the effect size?

Note: here Ho is drawn in blue, and H1 in red.



Distribution of male heights in North Korea and South Korea.

Sample 16 male heights from North Korea

mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

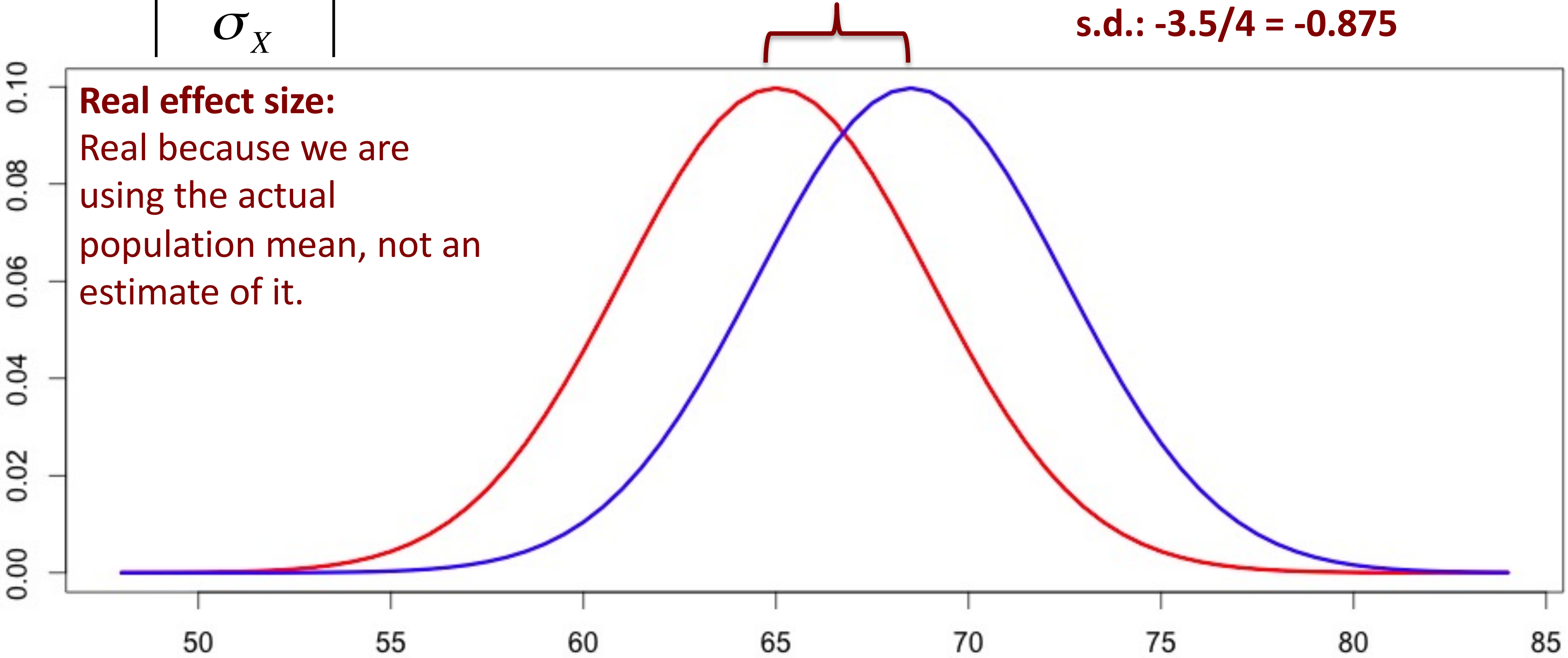
mean=68.5” sd=4”

What is the effect size?

$$d = \left| \frac{\mu_T - \mu_0}{\sigma_X} \right| = 0.875$$

Distance between means: -3.5

Distance between means in units of s.d.: -3.5/4 = -0.875



Real effect size:
Real because we are using the actual population mean, not an estimate of it.

Distribution of male heights in North Korea and South Korea.

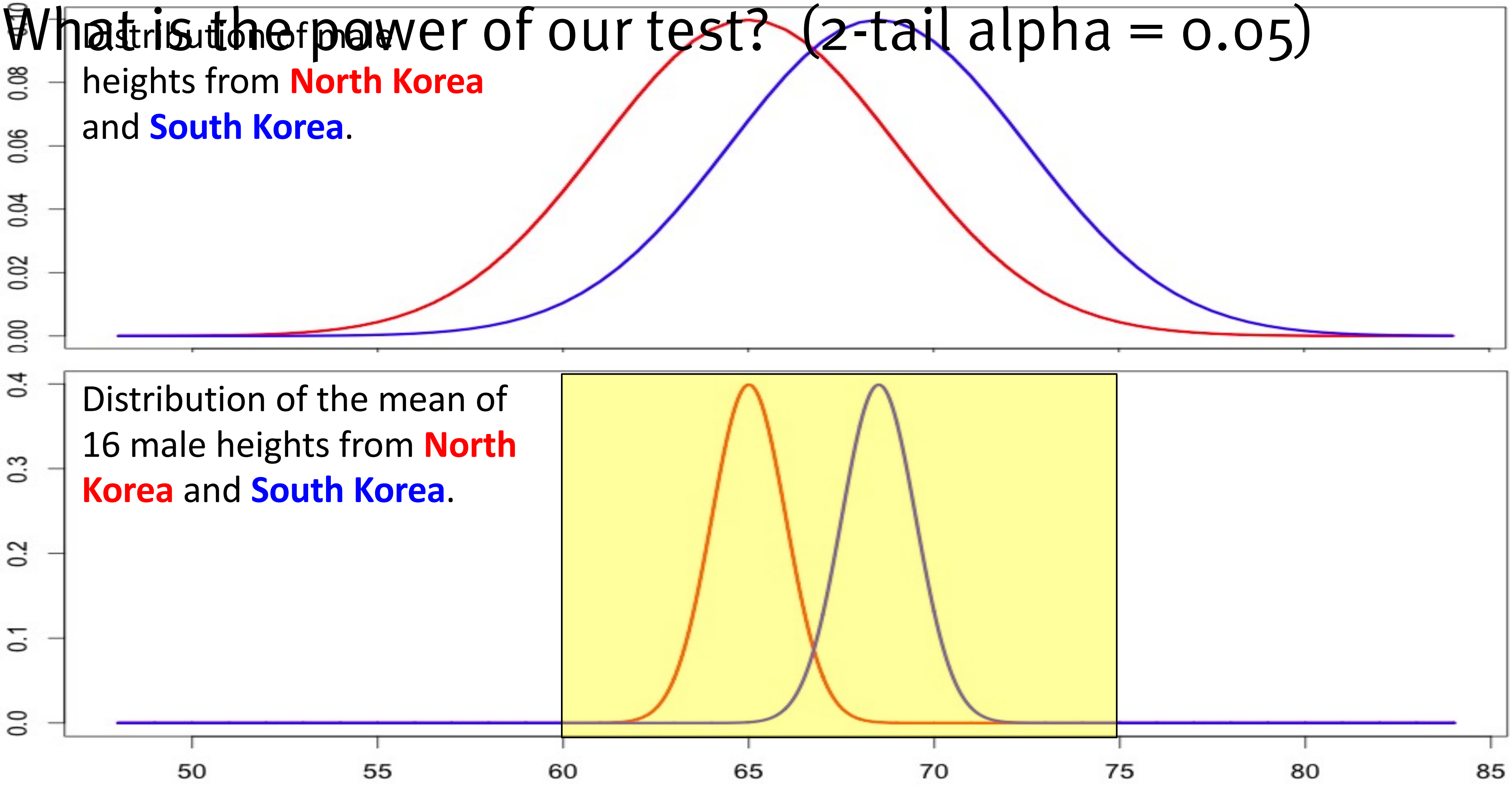
Sample 16 male heights from North Korea

mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5”   sd=4”

What is the power of our test? (2-tail alpha = 0.05)



Distribution of male heights from North Korea and South Korea.

Distribution of the mean of 16 male heights from North Korea and South Korea.

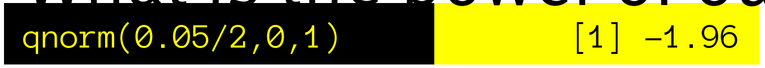
Sample 16 male heights from North Korea

mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5”   sd=4”

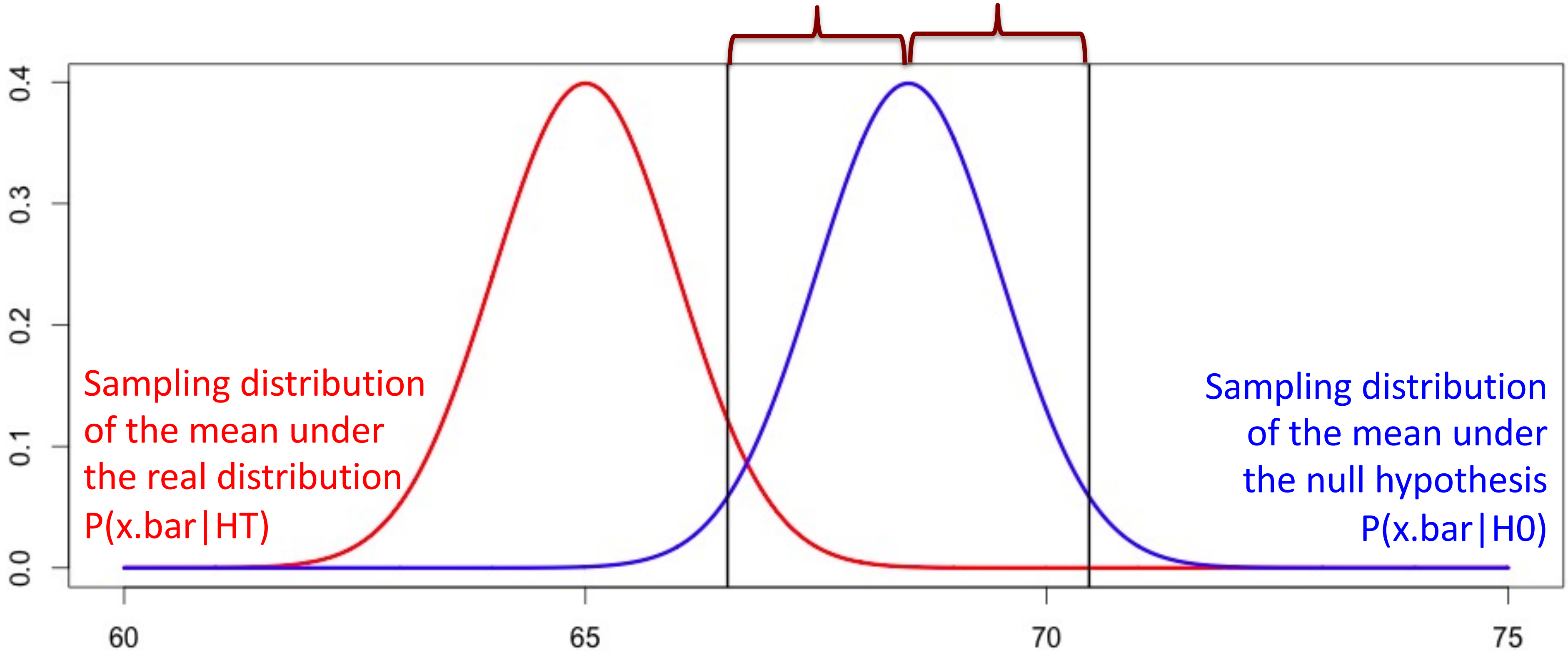
What is the power of our test? (2-tail alpha = 0.05)



```
qnorm(0.05/2,0,1)          [1] -1.96
```

Critical distance from H0 mean in s.e.m: 1.96

Critical distance from H0 mean in inches: 1.96*1



Distribution of the mean of 16 male heights from **North Korea** and **South Korea**.

Sample 16 male heights from North Korea

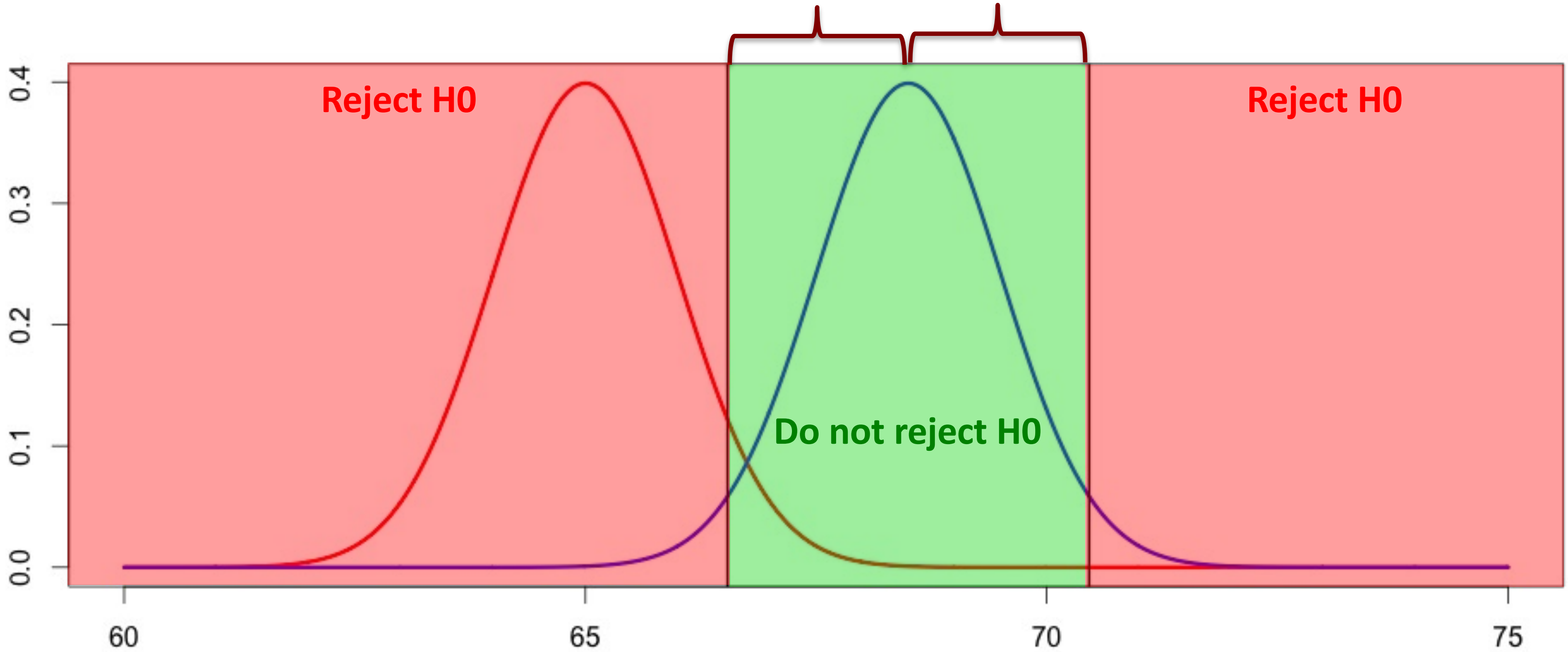
mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What is the power of our test? (2-tail alpha = 0.05)

Critical distance from H0 mean in std. errs. of the mean: 1.96



Distribution of the mean of 16 male heights from North Korea and South Korea.

Sample 16 male heights from North Korea

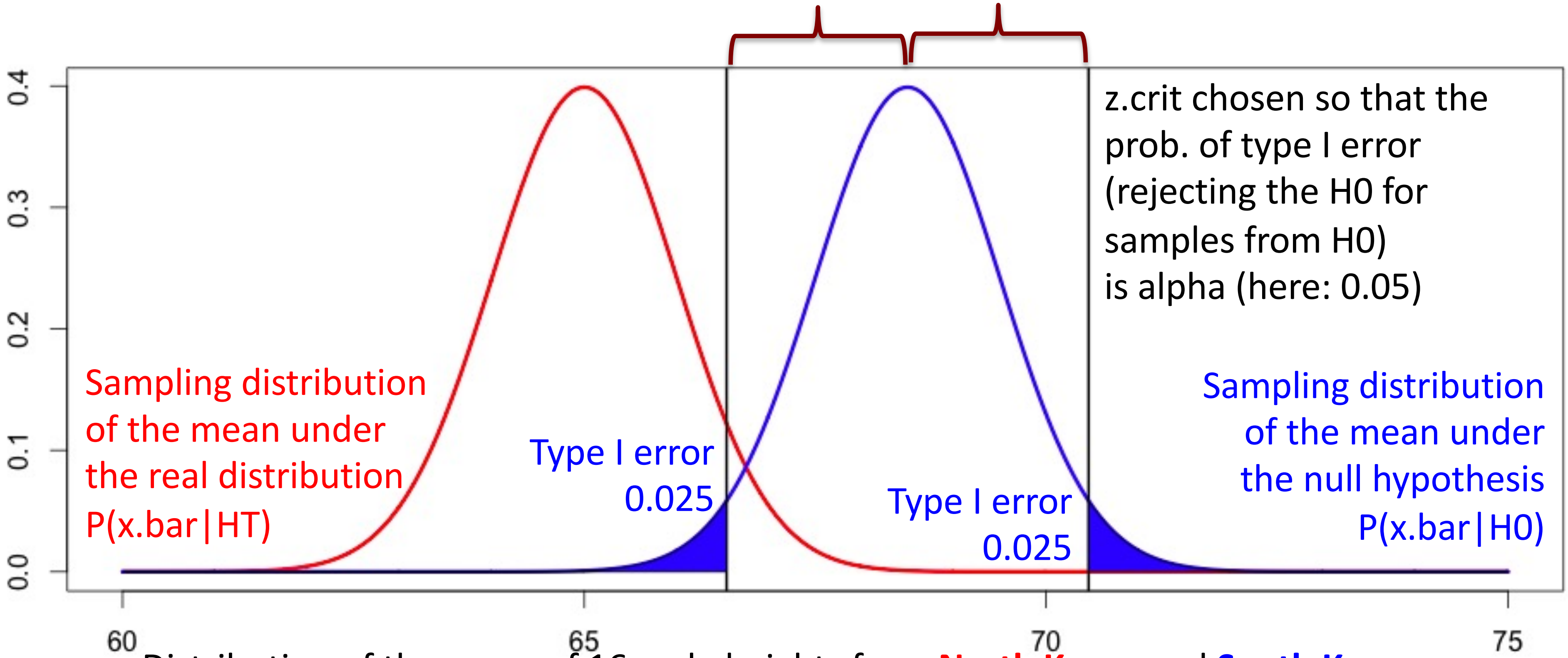
mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What is the power of our test? (2-tail alpha = 0.05)

Critical distance from H0 mean in std. errs. of the mean: 1.96



z.crit chosen so that the prob. of type I error (rejecting the H0 for samples from H0) is alpha (here: 0.05)

Sampling distribution of the mean under the real distribution P(x.bar|HT)

Type I error 0.025

Type I error 0.025

Sampling distribution of the mean under the null hypothesis P(x.bar|H0)

Distribution of the mean of 16 male heights from **North Korea** and **South Korea**.

Sample 16 male heights from North Korea

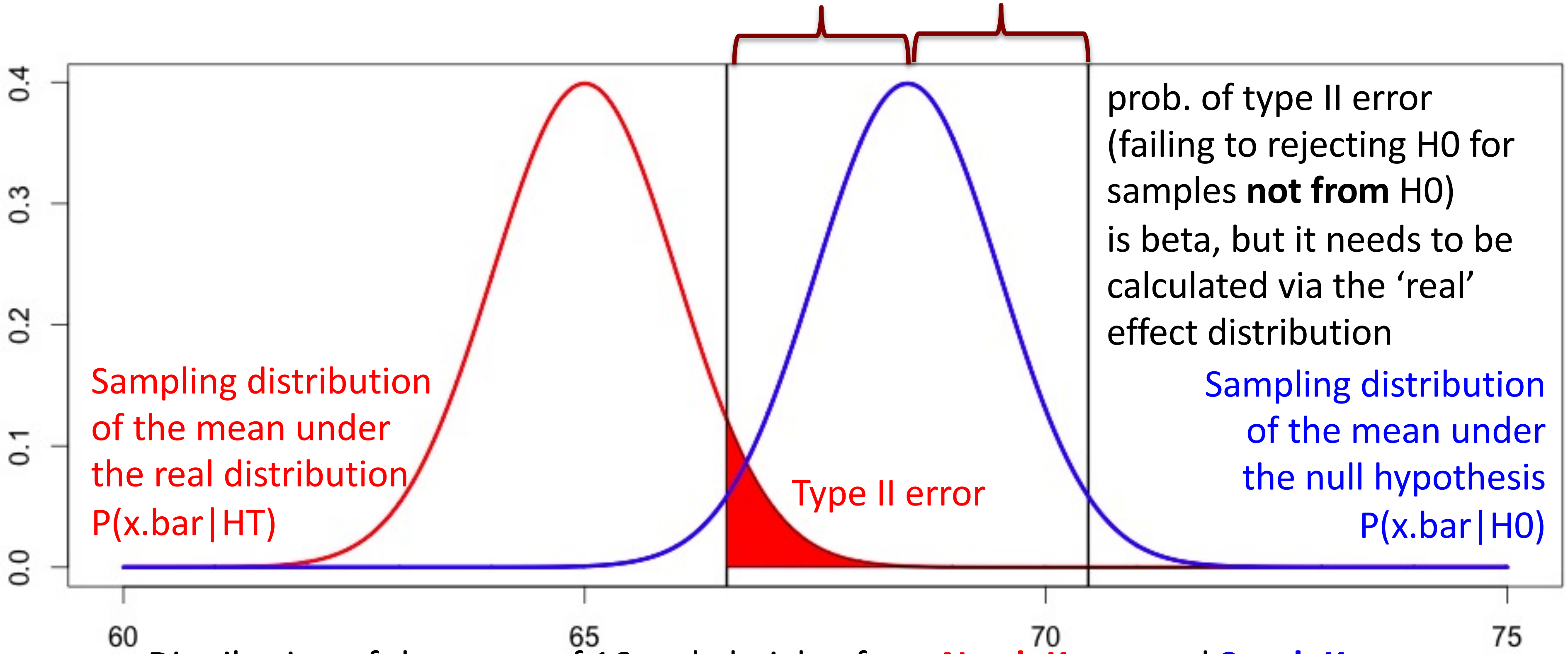
mean=65”   sd=4”

Test if mean male N.Korean heights are different from mean male S.Korean heights:

mean=68.5”sd=4”

What is the power of our test? (2-tail alpha = 0.05)

Critical distance from H0 mean in std. errs. of the mean: 1.96



prob. of type II error (failing to rejecting H0 for samples **not from** H0) is beta, but it needs to be calculated via the ‘real’ effect distribution

Sampling distribution of the mean under the real distribution P(x.bar|HT)

Sampling distribution of the mean under the null hypothesis P(x.bar|H0)

Distribution of the mean of 16 male heights from **North Korea** and **South Korea**.

Sample 16 male heights from North Korea

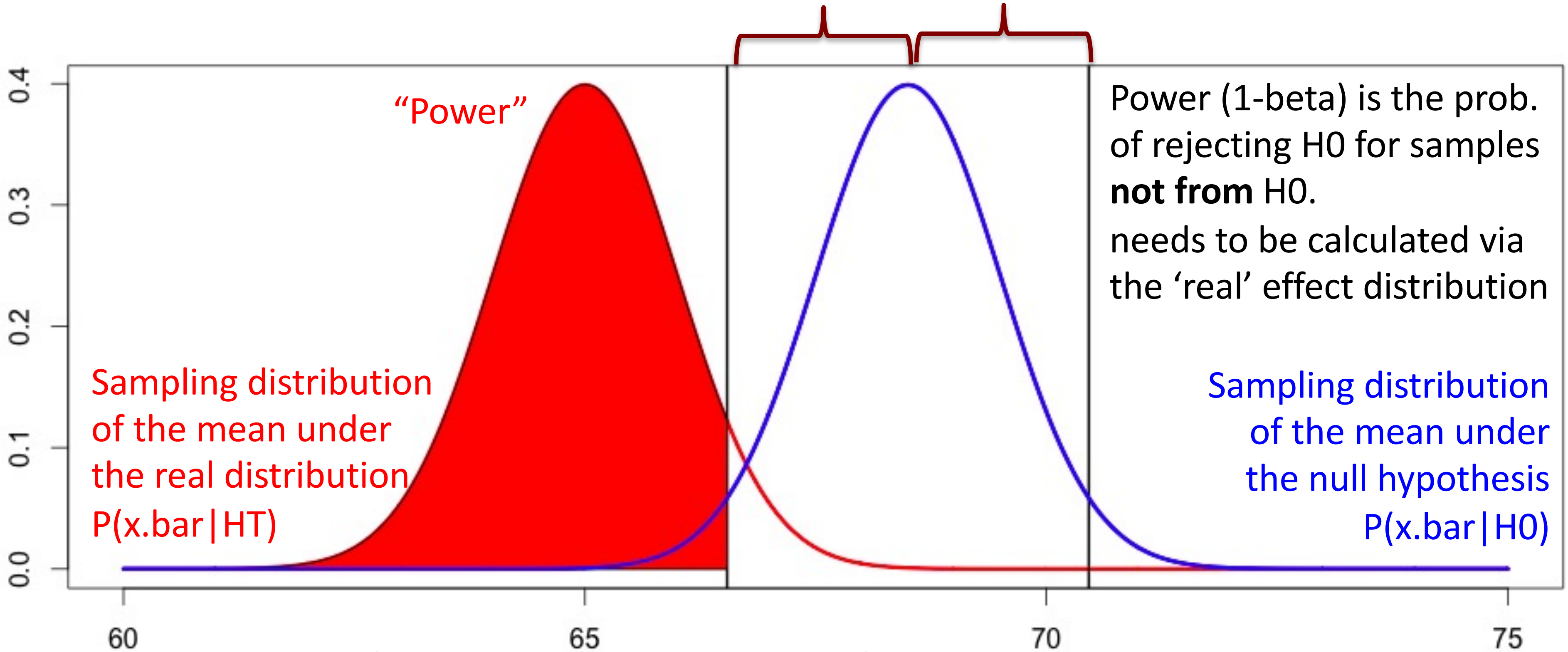
mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What is the power of our test? (2-tail alpha = 0.05)

Critical distance from H0 mean in std. errs. of the mean: 1.96



“Power”

Sampling distribution of the mean under the real distribution P(x.bar|HT)

Power (1-beta) is the prob. of rejecting H0 for samples **not from** H0.
needs to be calculated via the ‘real’ effect distribution

Sampling distribution of the mean under the null hypothesis P(x.bar|H0)

Distribution of the mean of 16 male heights from **North Korea** and **South Korea**.

Sample 16 male heights from North Korea

mean=65” sd=4”

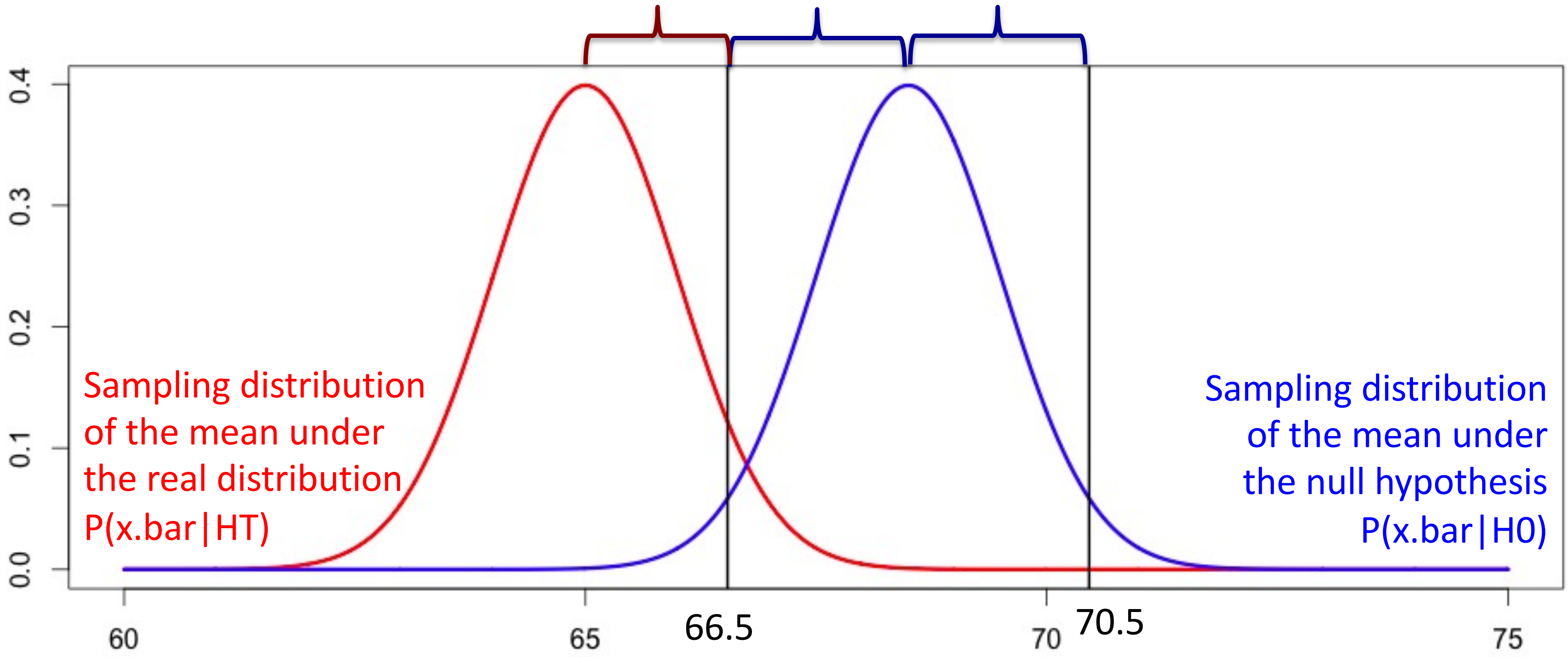
Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What is the power of our test? (2-tail alpha = 0.05)

Z crit rel. to H_T: 1.54

Z crit rel. to H_0: 1.96



Sampling distribution of the mean under the real distribution P(x.bar|HT)

Sampling distribution of the mean under the null hypothesis P(x.bar|H0)

Distribution of the mean of 16 male heights from **North Korea** and **South Korea**.

Sample 16 male heights from North Korea

mean=65”   sd=4”

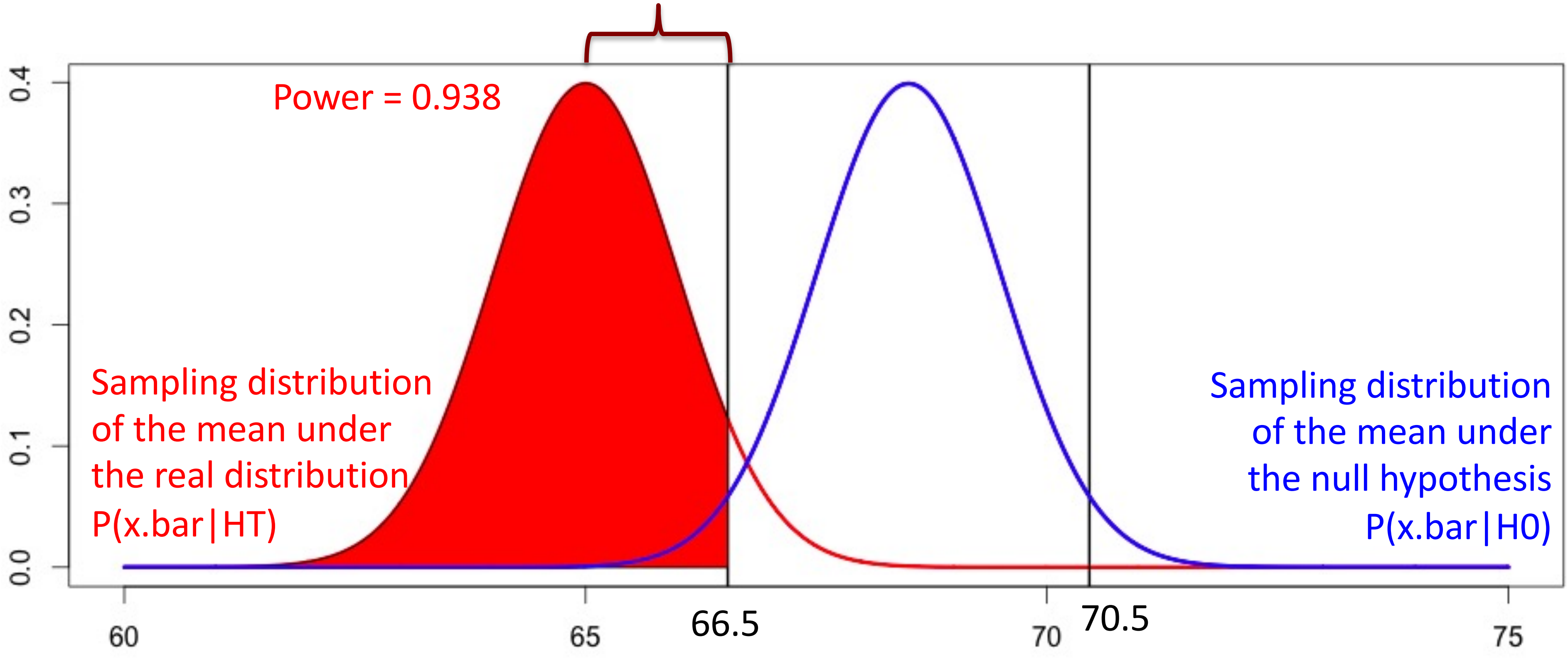
Test if sample mean is different from male S.Korean heights:

mean=68.5”   sd=4”

What is the power of our test? (2-tail alpha = 0.05)

```
pnorm(1.54,0,1)
[1] 0.938
```

Z.crit rel. to HT: 1.54



Distribution of the mean of 16 male heights from **North Korea** and **South Korea**.

Sample 16 male heights from North Korea

mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

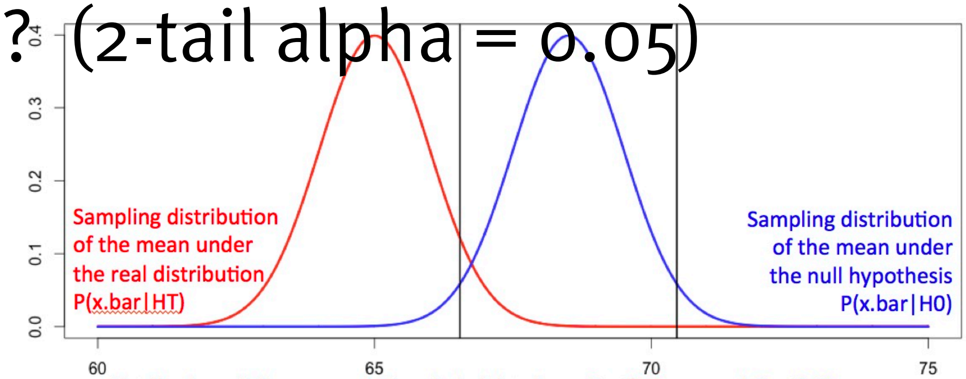
mean=68.5”   sd=4”

What is the power of our test? (2-tail alpha = 0.05)

We could have done this much faster!

Option 2: We have the true means, and standard deviations, so we don’t need to mess with z-scores!

We also see that the true distribution is below the H0 distribution…



Distribution of the mean of 16 male heights from North Korea and South Korea.

```
qnorm(0.05/2,68.5,4/sqrt(16))
```

```
[1] 66.54
```

Lower critical x.bar

```
pnorm(66.54,65,4/sqrt(16))
```

```
[1] 0.938
```

Proportion of true dist. below that critical x.bar. (power!)

Be careful of the signs: Upper or lower tail from null? Upper or lower tail from real?

Sample 16 male heights from North Korea

mean=65”   sd=4”

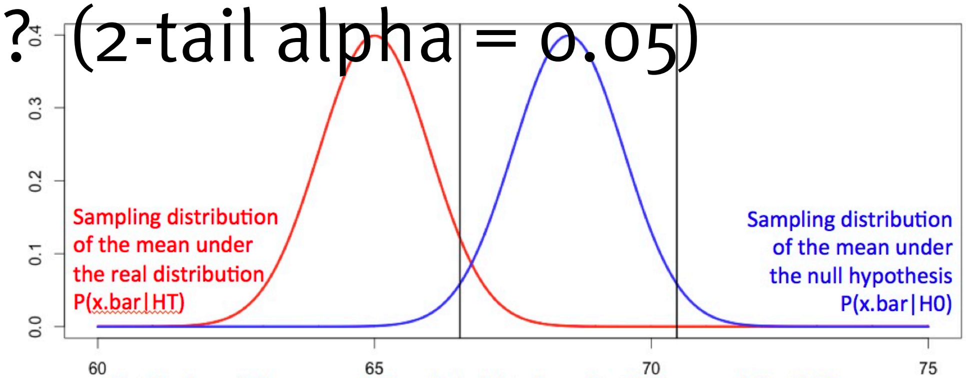
Test if sample mean is different from male S.Korean heights:

mean=68.5”   sd=4”

What is the power of our test? (2-tail alpha = 0.05)

We could have done this much faster!

Option 3: Using slick math, relying on equal variance, and symmetry of the Normal distribution, we could use the z-scores and effect sizes without ever translating into x.bar.



Distribution of the mean of 16 male heights from North Korea and South Korea.

```
abs(qnorm(0.05/2,0,1)                    [1] 1.96
```

Critical z.score rel. to H0 (positive)

```
1.96-d*sqrt(n)
1.96-0.875*sqrt(16)                      [1] -1.54
```

Critical z.score rel to HT (having flipped it around to be bigger than H0!)

$$d = \left|\frac{\mu_T - \mu_0}{\sigma_X}\right| = 0.875$$

```
1-pnorm(-1.54,0,1)                       [1] 0.938
```

Power!

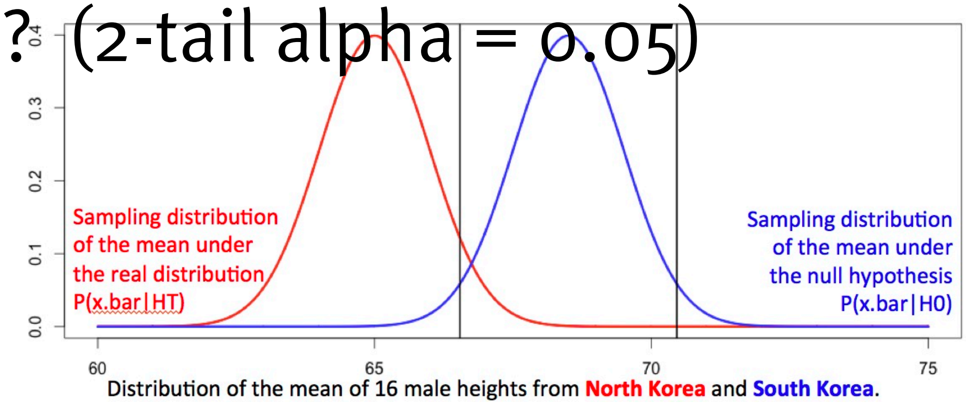
Sample 16 male heights from North Korea

mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What is the power of our test? (2-tail alpha = 0.05)



Option 1: Obtain critical z-scores relative to H0, convert into critical x.bar, convert into z-score relative to HT, calculate power.
(sort of stupid and slow, but makes the process explicit)

Option 2: Obtain critical x.bar, calculate power.
(you need to know true/null means and s.d.s – can’t work with effect sizes)

Option 3: Obtain critical z-score rel to H0, use power to calculate z-score relative to HT, calculate power.
(probably the most general/useful formulation)

```
1-pnorm(abs(qnorm(a/2))-d*sqrt(n))
```

General note: Make sure you consider what the distributions look like! (sample from world?)

Sample 16 male heights from North Korea

mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5”   sd=4”

What is the power of our test? (2-tail alpha = 0.05)

```
pwr::pwr.norm.test(d,n)
```

```
pwr::pwr.norm.test(0.875,16)
                                                 0.938
```

Sample *n* male heights from North Korea

mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)

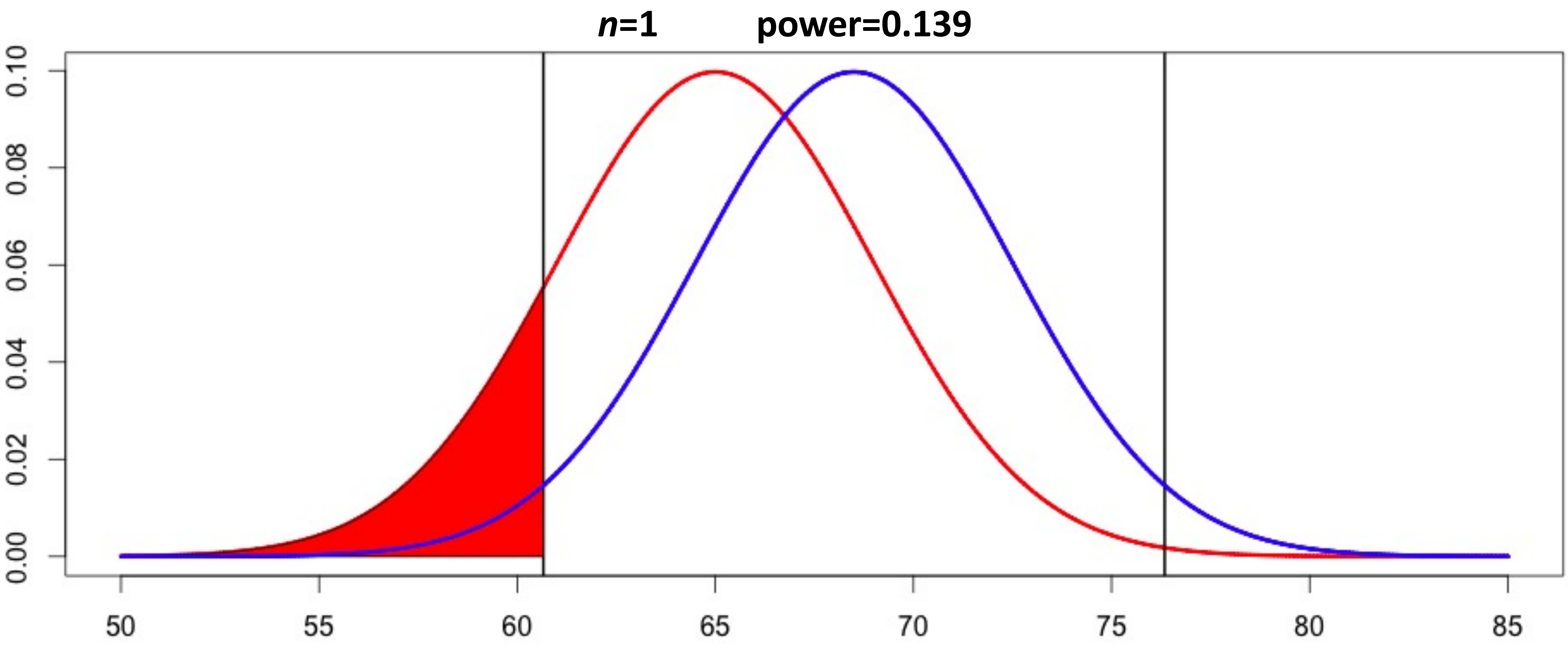
Sample *n* male heights from North Korea

mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)



Distribution of male heights in North Korea and South Korea.

Sample *n* male heights from North Korea

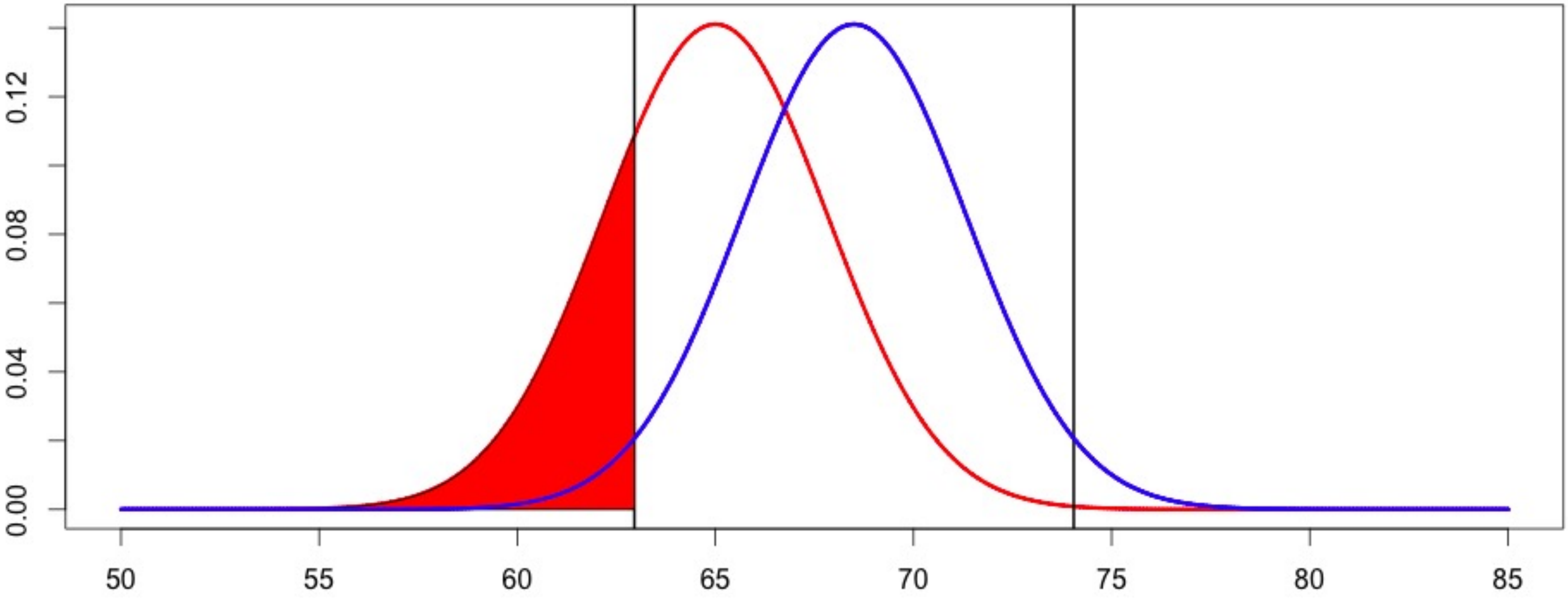
mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)

n=2 power=0.235



Distribution of male heights in North Korea and South Korea.

Sample *n* male heights from North Korea

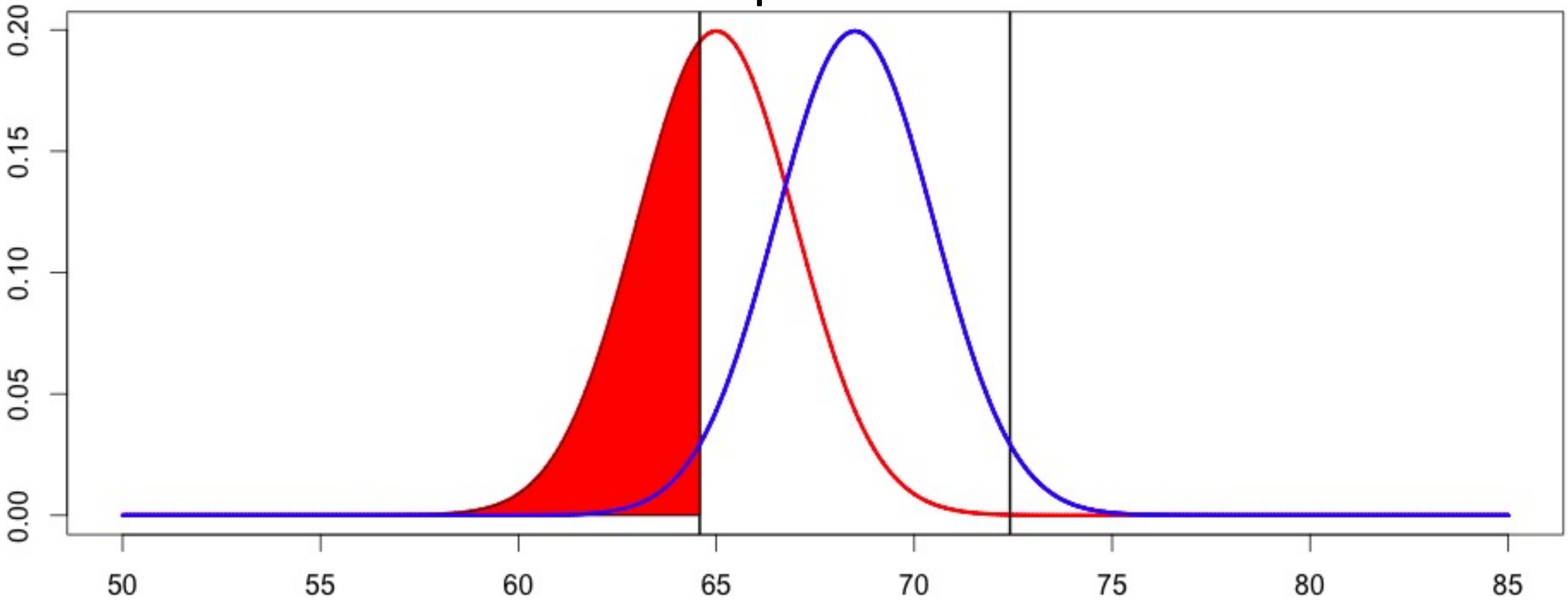
mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)

n=4   power=0.417



Distribution of male heights in North Korea and South Korea.

Sample *n* male heights from North Korea

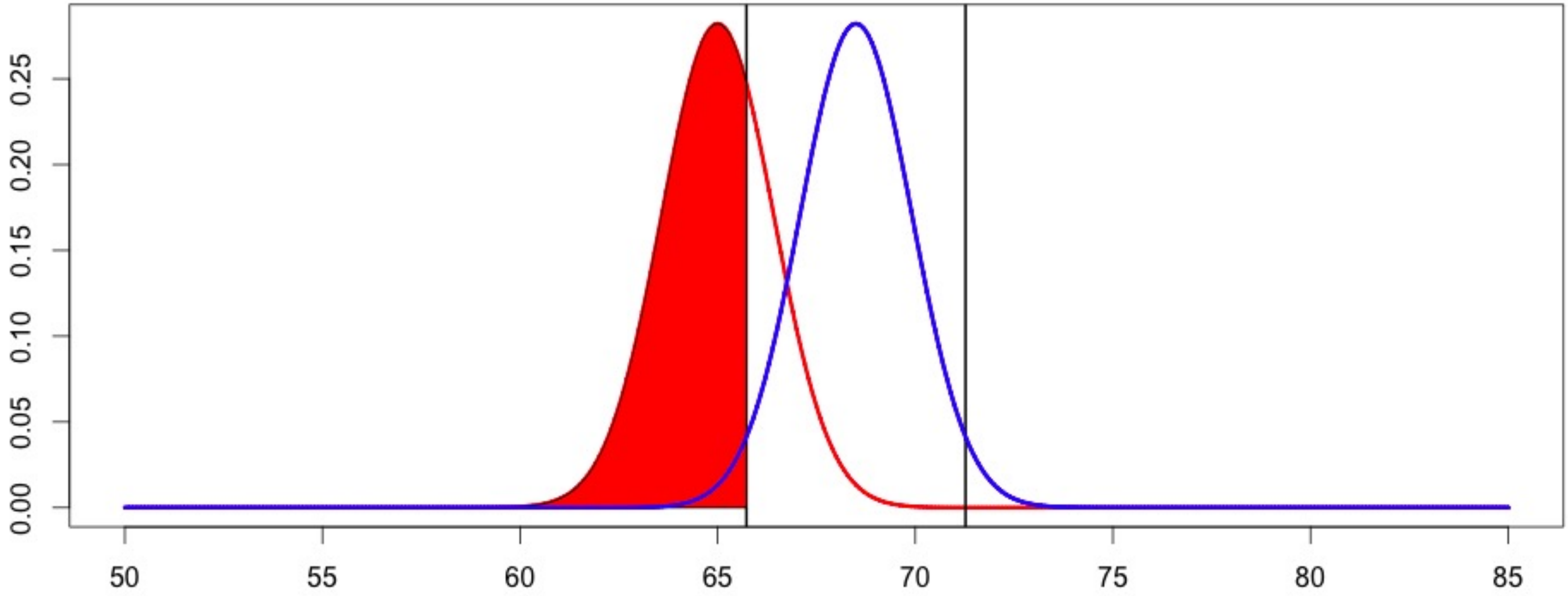
mean=65” sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5”sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)

***n*=8 power=0.697**



Distribution of male heights in North Korea and South Korea.

Sample *n* male heights from North Korea

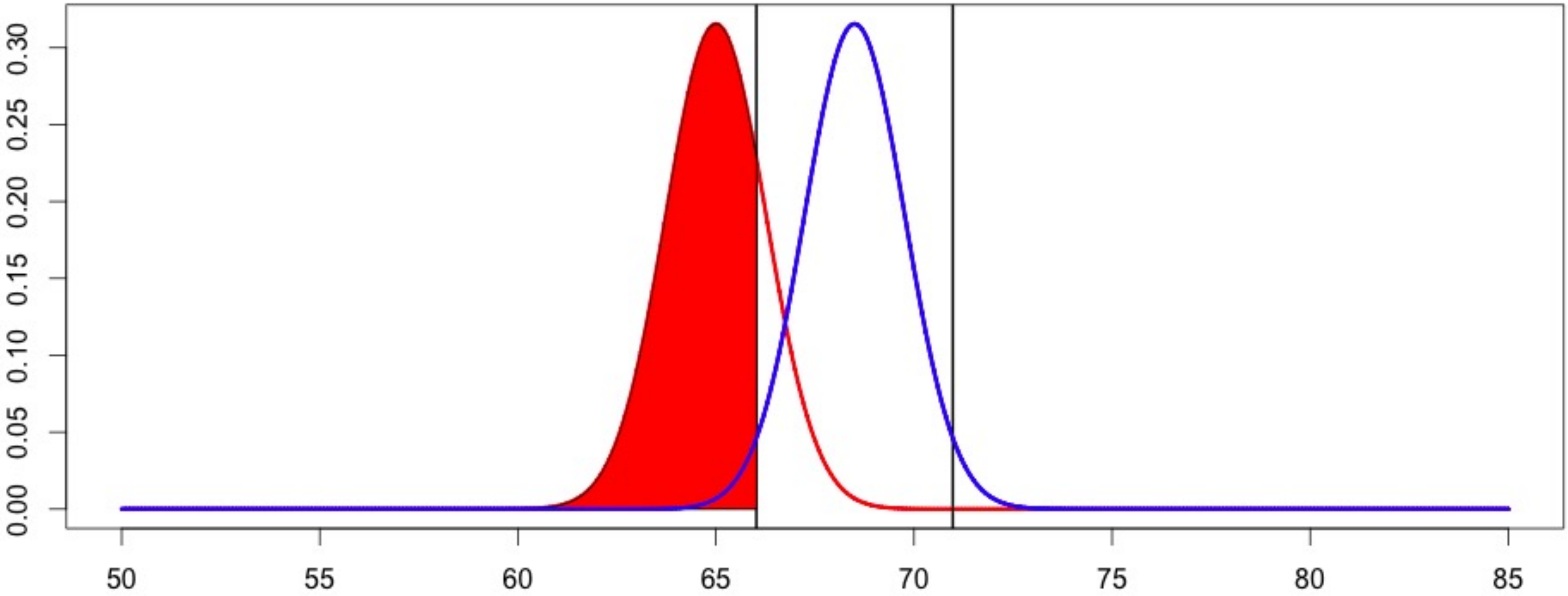
mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)

***n*=10   power=0.79**



Distribution of male heights in **North Korea** and **South Korea**.

Sample *n* male heights from North Korea

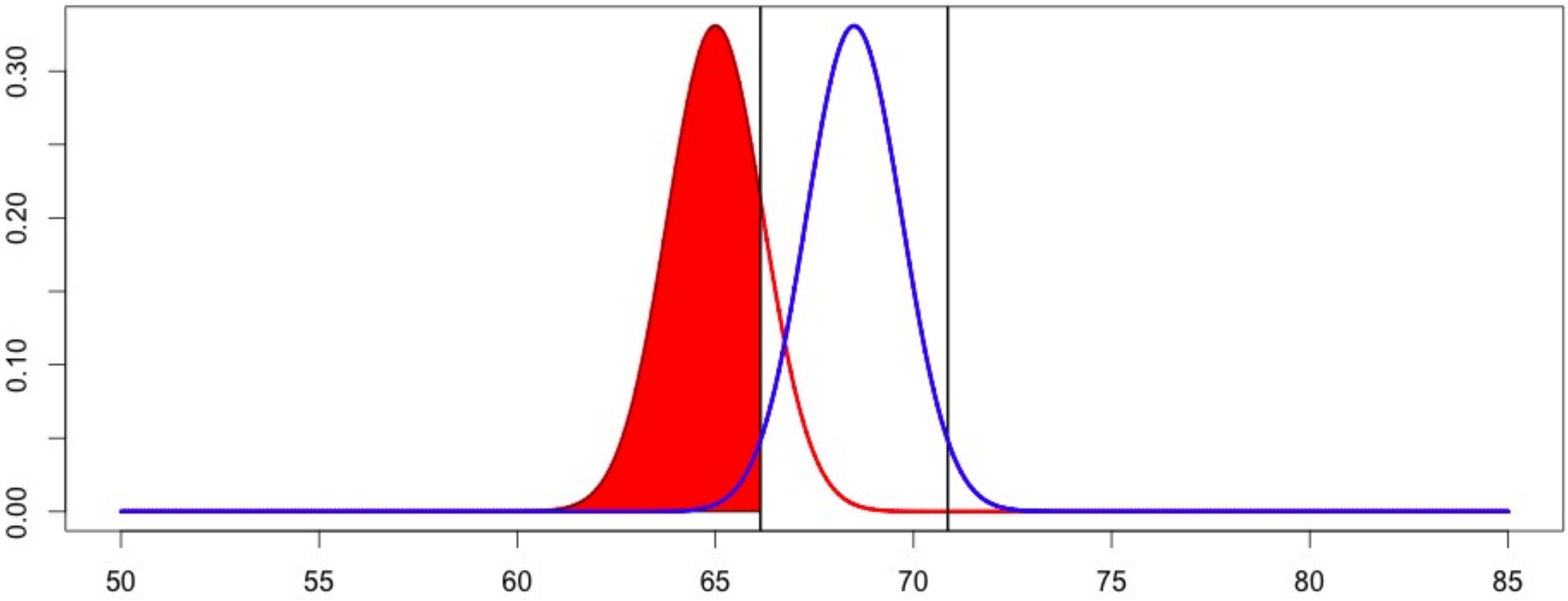
mean=65”   sd=4”

Test if sample mean is different from male S.Korean heights:

mean=68.5” sd=4”

What should *n* be for our power to be 0.8? (2-tail alpha = 0.05)

_n_=11   power=0.827



Distribution of male heights in **North Korea** and **South Korea**.

Z-test power functions

- Get the power given *d*, *n*, and *alpha*. (2-tailed!)

```
pwr::pwr.norm.test(d=d, n=n, sig.level=alpha)
```

- Get the necessary n to reach *power,* given *d,* and *alpha.*

```
pwr::pwr.norm.test(d=d, sig.level=alpha, power=power)
```

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals
- Null hypothesis significance testing
- Power

q=(1-α)% confidence interval on mean

$$\bar{x} \pm z_{\alpha/2}\sigma_0 / \sqrt{n}$$

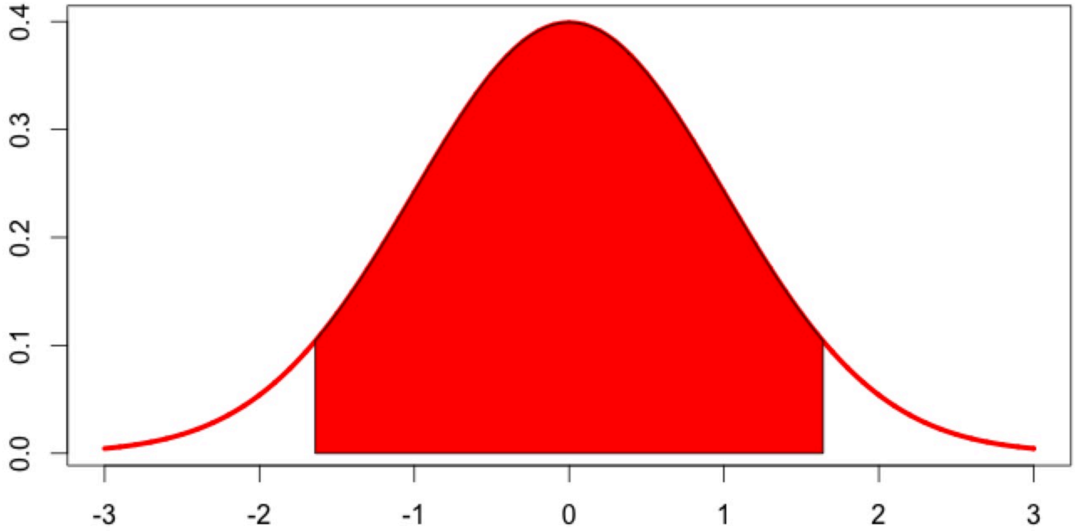
Critical z score?

- What is the Z_crit such that q% of of all z-scores are between –Z_crit and +Z_crit? E.g., q=90%
- What is the distance between the sample mean and the population mean (in units of standard errors of the mean) such 90% of the other potentially sampled distances are less than this one?

```
Q                                              [1] 0.9
(1-Q)/2                                        [1] 0.05
```

So 90% in the middle of the dist. Therefore 5% in either tail.

```
Z_crit = qnorm((1-Q)/2)                        [1] -1.64
```



90% of Z-scores are within 1.64 of 0.0.

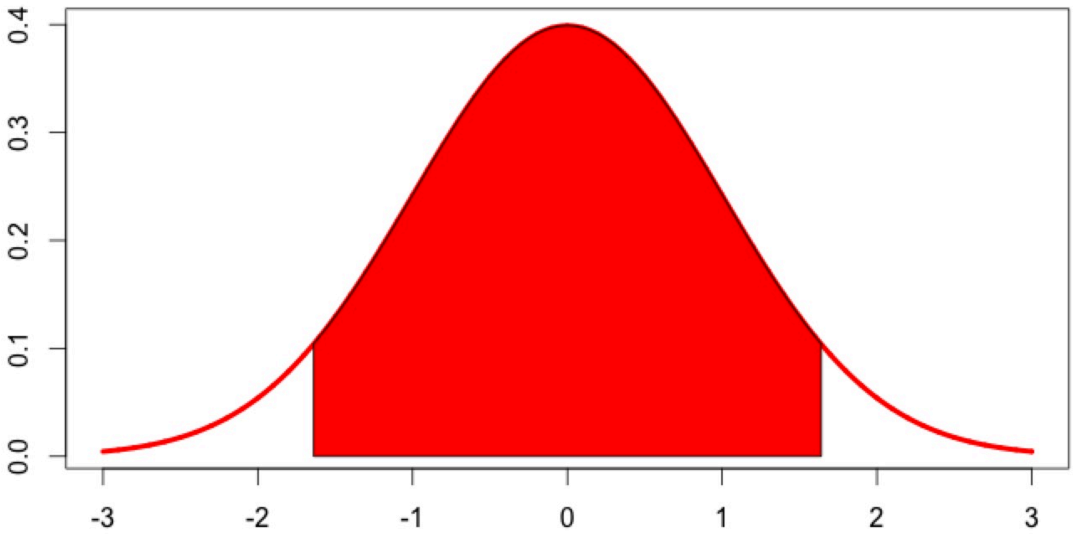
90% of sample means are within 1.64 s.e.m of the population mean.

A confidence interval

We have: a sample mean, n, population sd, and get s.e.m.

```
x_bar = mean(x)                          [1] 108
n = length(x)                             [1] 16
sdX = 15
sem = sdX/sqrt(n)                        [1] 3.75

Z_crit = qnorm((1-Q)/2)                 [1] -1.64
```



90% of Z-scores are within 1.64 of 0.

90% of sample means are within 1.64 s.e.m.s of population mean.

90% of sample means are within 1.64*3.75=6.15 IQ points of the population mean

90% interval on population mean
108-6.15 to 108+6.15
[101.85 to 114.15]

Confidence interval on pop. mean

- Q% confidence interval:
  Sample mean +/- z_crit * sem
  z_crit defined such P(abs(z) < z_crit) = Q%

```
Z_crit = qnorm((1-Q)/2)                                  [1] -1.64
```

90% z-score interval on deviation of sample mean from population mean (in standard errors of the mean):

[-1.64 to 1.64]

90% interval on deviation of sample mean from population mean (in units of x, here IQ):

[-1.64*sem to 1.64*sem] = [-6.15 to 6.15]

90% interval on population mean:

x_bar + [-1.64*sem to 1.64*sem] = [101.85 to 114.15]

Confidence interval on pop. mean

- Q% confidence interval:
  Sample mean +/- z_crit * sem
  z_crit defined such P(abs(z) ‹ z_crit) = Q%

```
Z_crit = qnorm((1-Q)/2)
```

```
[1] -1.64
```

$$\bar{x} \pm z_{\alpha/2}\sigma_0 / \sqrt{n}$$

Mean(x) = 110. n=25. st.dev = 15.

What is the 75% (z-score) confidence interval on the mean?

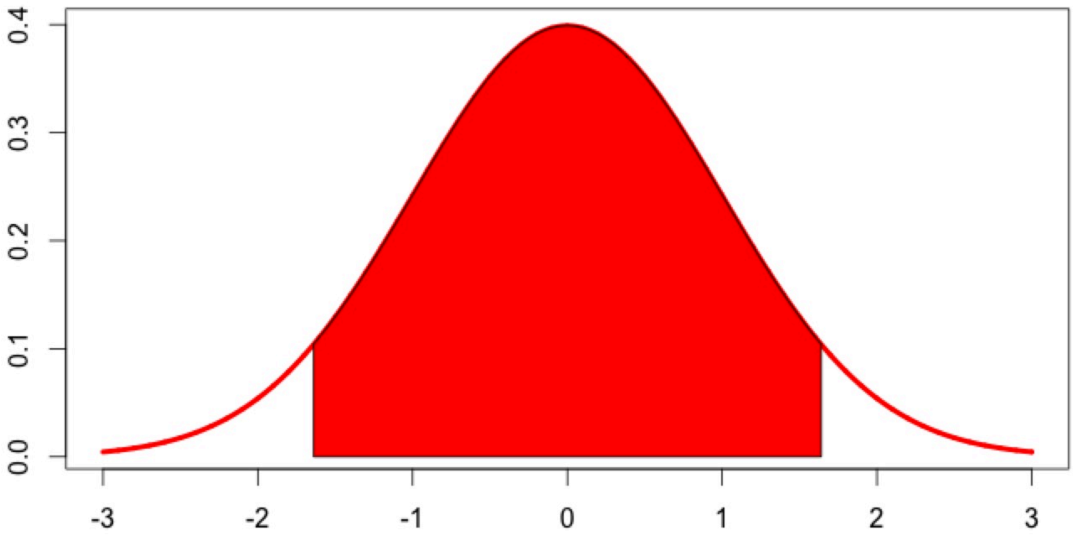
Let’s calculate a confidence interval.

```
x                                   [16] 120 113 129 113 …
```

- What if we think these folks might have different mean and a *different sd* than the overall IQ population… Can’t we just use the sample sd to define sem?
  - Not with the z (normal) distribution.
  - We will use the “T” distribution. More later

  (This is why Z-score confidence intervals are good for illustration, but very rarely used in practice)

Confidence intervals are weird.



These are probability statements about the distribution of all possible sample means

90% of Z-scores are within 1.64 of 0.

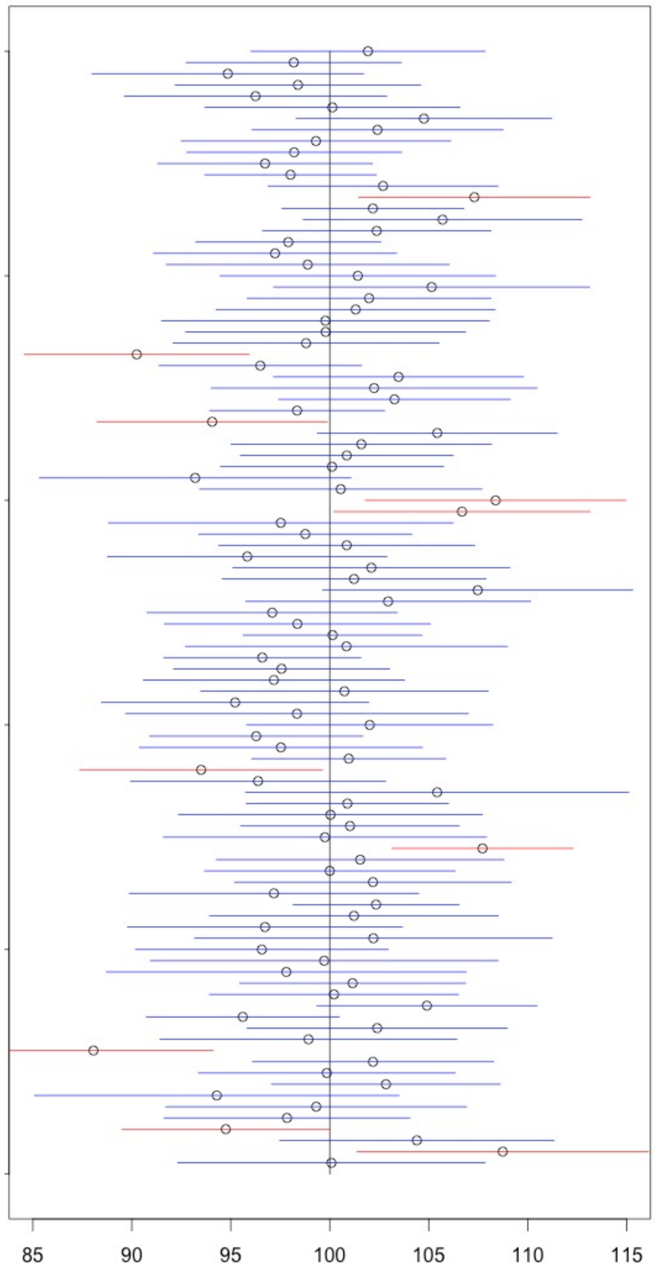
90% of sample means are within 1.64 s.e.m.s of population mean.

90% of sample means are within 1.64*3.75=6.15 inches of the population mean

How did we get to this statement about the population mean?

90% interval on population mean
108-6.15 to 108+6.15 = [101.85 to 114.15]

Confidence intervals are weird.



Sample 16 male heights. mean=64”.
Ho: Sample from South Korean male population:
mean=68.5” sd=4”

What’s our 95% confidence interval on the mean of this sample’s population?

Step 1) assume that this sample’s population has the same standard deviation in height as the population of South Koreans (so we can do a Z-test conf. interval). **<- ! This should worry you !**

Critical alpha

```
(1-0.95)                                [1] 0.05
```

Critical Z

```
abs(qnorm(0.05/2,0,1))                  [1] 1.96
```

CI lower bound

```
64-1.96*(4/sqrt(16))                   [1] 62.04
```

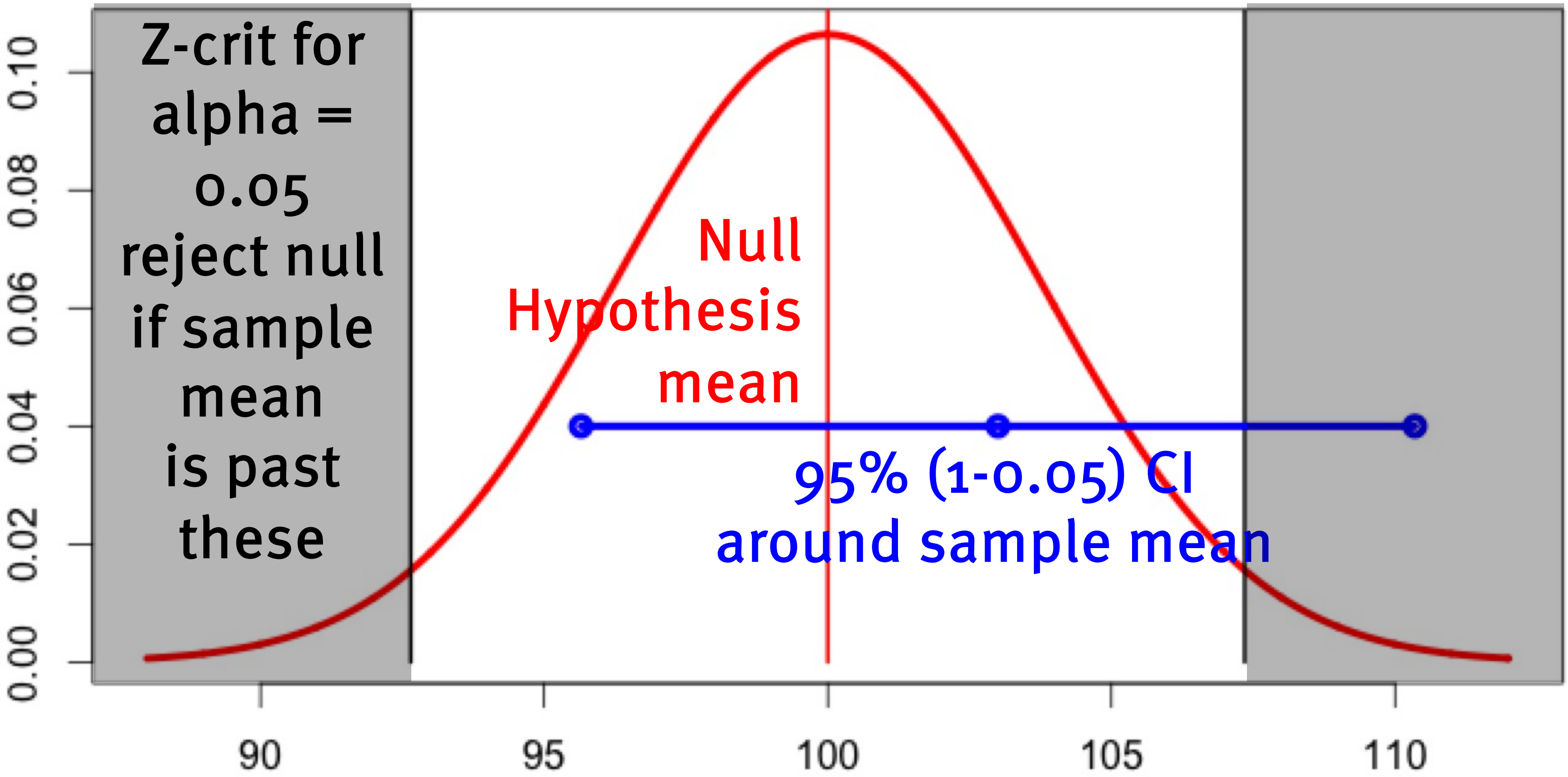
CI upper bound

```
64+1.96*(4/sqrt(16))                   [1] 65.96
```

So our 95% confidence interval on the mean is [62, 66]

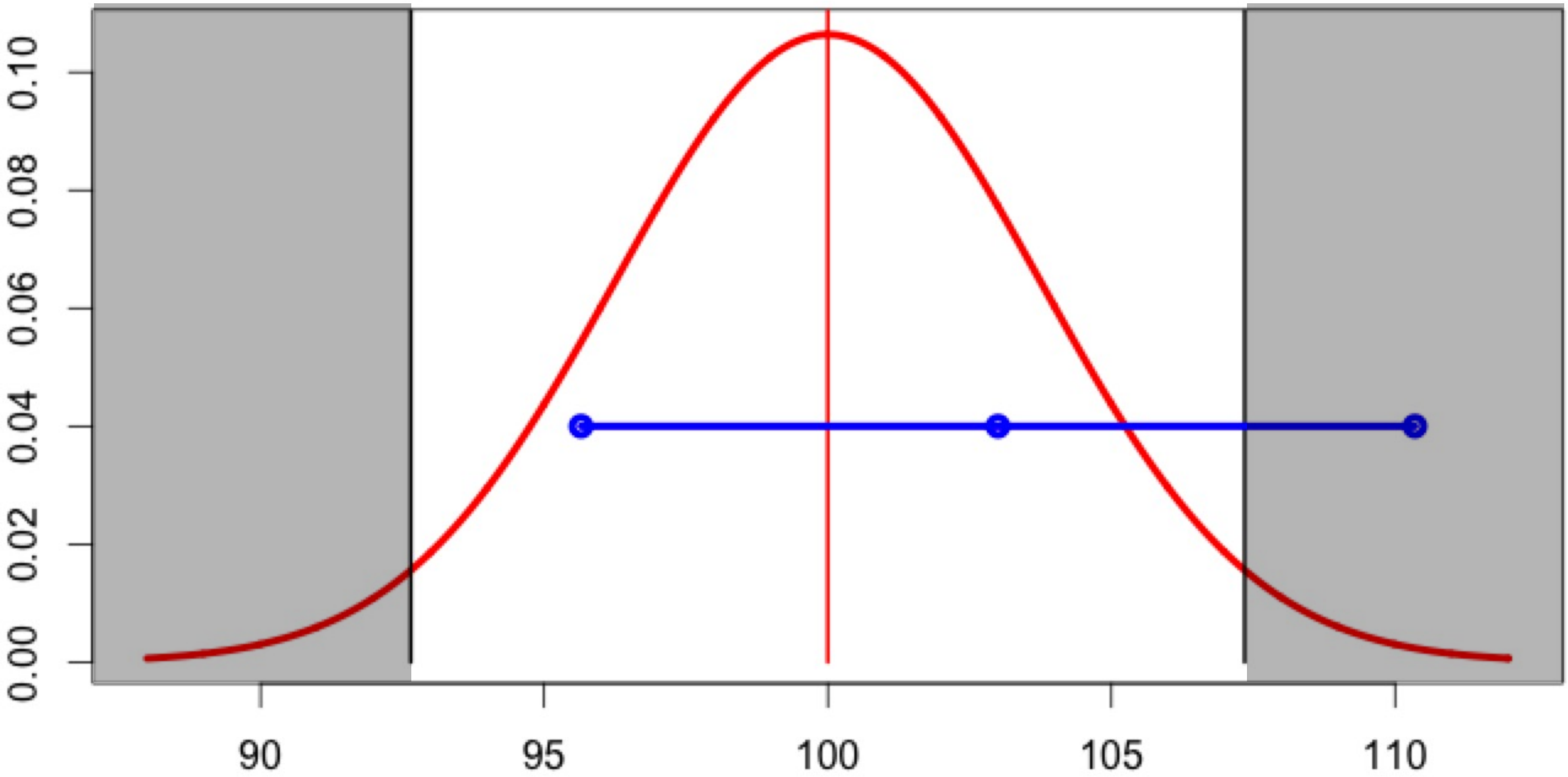
Confidence intervals and NHSTs on mean

- We reject the null hypothesis at a certain alpha if…
  - The z.score is larger (absolute value) than z.crit (ergo p ‹ α)
  - The sample mean is further than z.crit*sem from null mean
  - The (1-α) confidence interval excludes the null mean



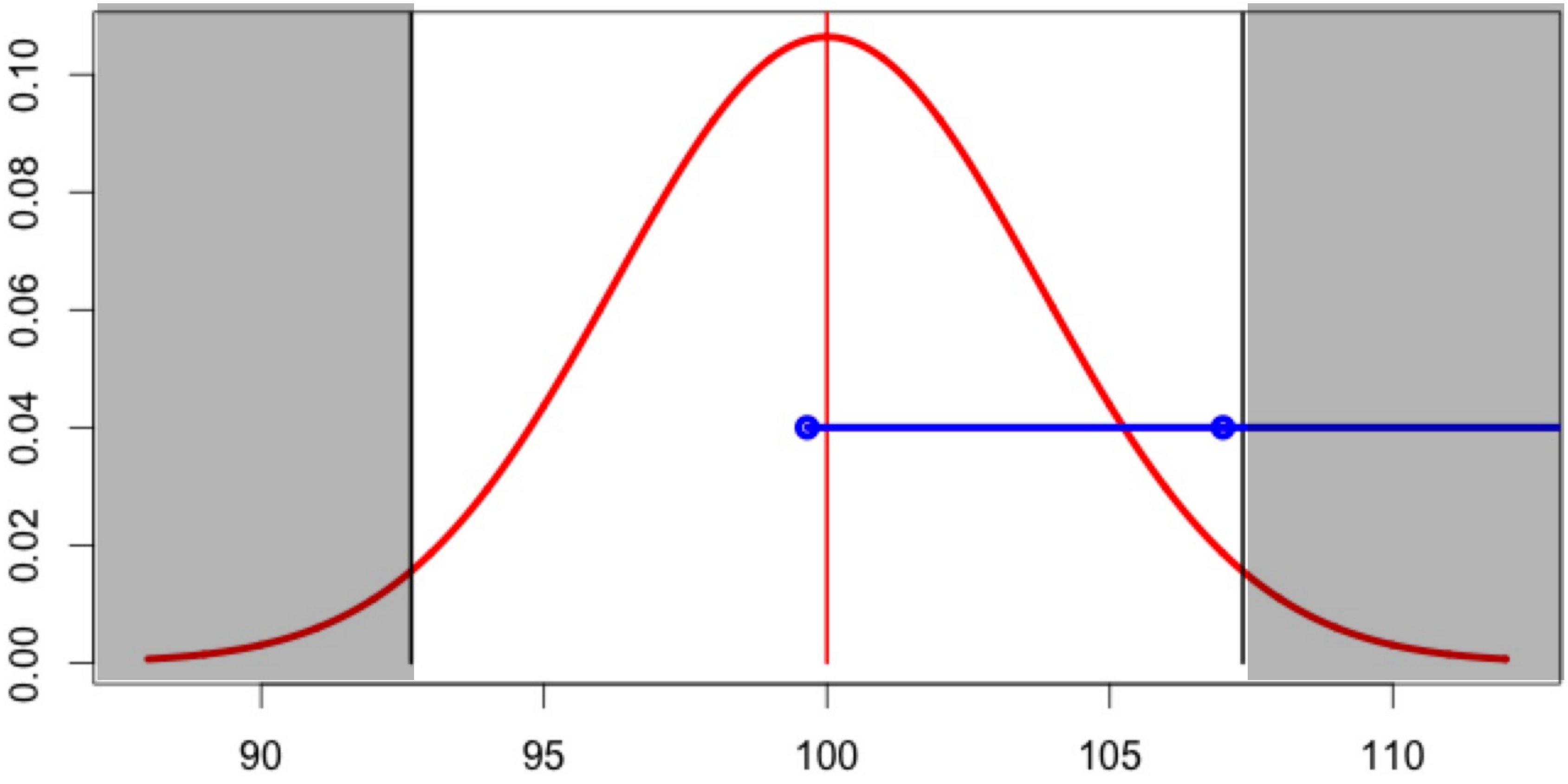
Confidence intervals and NHSTs on mean

- We reject the null hypothesis at a certain alpha if…
  - The z.score is larger (absolute value) than z.crit (ergo $p < \alpha$)
  - The sample mean is further than z.crit*sem from null mean
  - The ($1-\alpha$) confidence interval excludes the null mean



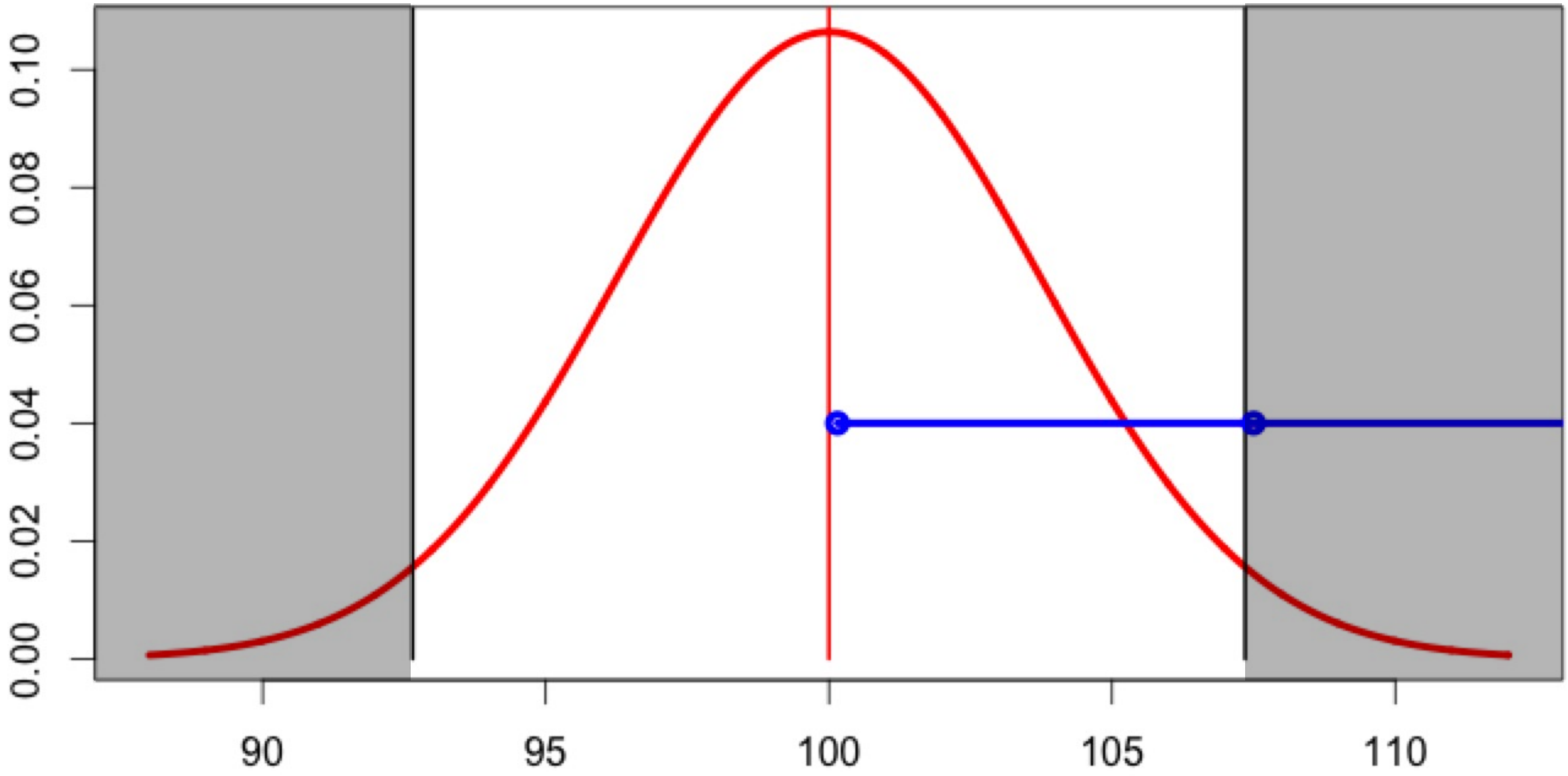
Confidence intervals and NHSTs on mean

- We reject the null hypothesis at a certain alpha if…
  - The z.score is larger (absolute value) than z.crit (ergo p ‹ α)
  - The sample mean is further than z.crit*sem from null mean
  - The (1-α) confidence interval excludes the null mean



Confidence intervals and NHSTs on mean

- We reject the null hypothesis at a certain alpha if…
  - The z.score is larger (absolute value) than z.crit (ergo p ‹ α)
  - The sample mean is further than z.crit*sem from null mean
  - The (1-α) confidence interval excludes the null mean



Confidence intervals and NHSTs on mean

- We reject the null hypothesis at a certain alpha if…
  - The z.score is larger (absolute value) than z.crit (ergo $p < \alpha$)
  - The sample mean is further than z.crit*sem from null mean
  - The $(1-\alpha)$ confidence interval excludes the null mean

- If the sample mean (x.bar) passes the critical rejection value (null mean ± z.crit*sem) then the null mean will fall outside the (x.bar ± z.crit*sem) confidence interval around the sample mean.
- If we can reject at alpha, then the null mean falls outside the 1-alpha confidence interval. And vice versa.

1-alpha confidence interval provides same NHST, but is more useful.

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals and NHST analogue
- Null hypothesis significance testing
- Power
- Wrap up

Normal variable stats.

- NHST: Z-test.
  - Get (2-tailed) p-value via

$$z_{\overline{x}} = \frac{\overline{x} - \mu_0}{\sigma_X}\sqrt{n}$$

```
2*pnorm(-abs(z),0,1)
```

- Confidence intervals on mean
  - Equivalent to null hypotheses!

$$\overline{x} \pm z_{\alpha/2}\sigma_0 / \sqrt{n}$$

```
za = qnorm(a/2,0,1)
```

- Effect size
  - Scale and sample size neutral.

$$d = \left|\frac{\mu_T - \mu_0}{\sigma_X}\right|$$

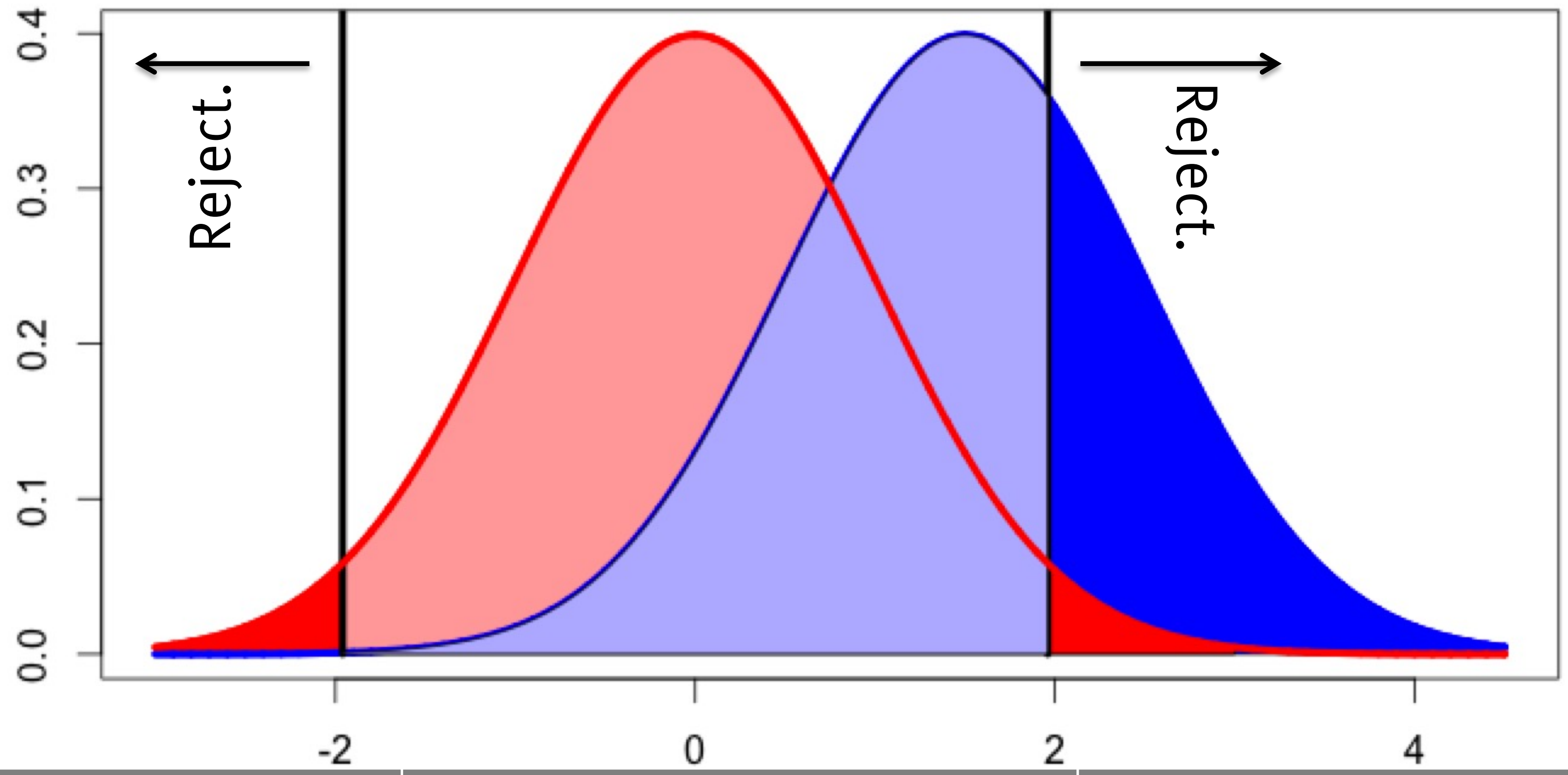
- Alpha, Beta, Power.
  - Effect size and n matter.

```
pow    =   pwr::pwr.norm.test(d, n, alpha)
```

```
n.needed    =  pwr::pwr.norm.test(d, power=power, alpha)
```

- Critically: We assert that we **know** the real standard deviation. We usually do not

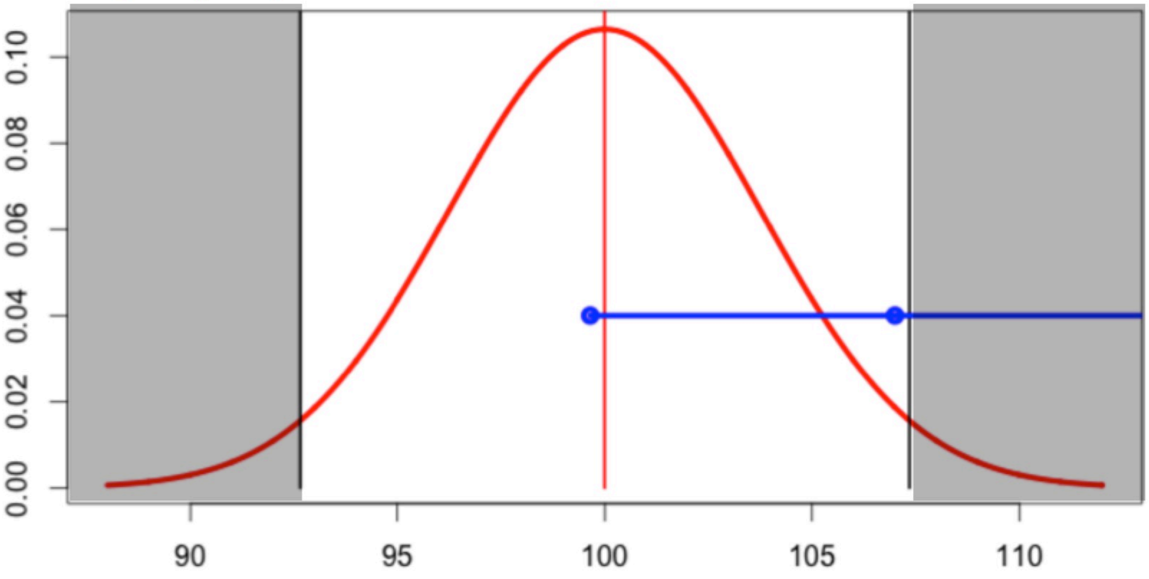
Errors in NHST

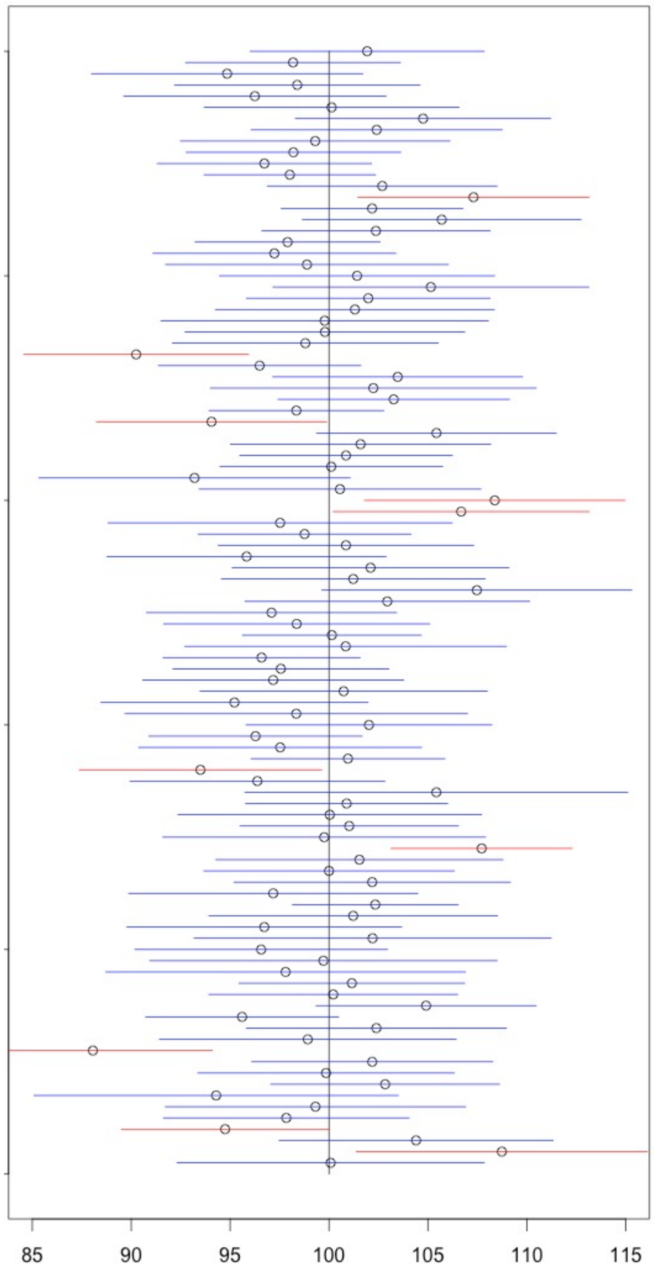


	H_0 false	H_0 true
Reject H_0	Correct rejection of null (Pr = 1-β 'power')	Type I error (Pr = α)
Fail to reject H_0	Type II error (Pr = β)	Correct failure to reject null (Pr = 1-α)

Confidence intervals

- If a 90% confidence interval on the mean excludes the null hypothesis mean, we can reject that null hypothesis with 2-tailed alpha = 0.1, and vice versa.





- We expected 90 out of 100 90% confidence intervals to include the true mean.

  "90%" refers to a long-run property of the procedure used to define the confidence interval, not to the specific confidence interval you have.

Probabilities in classical statistics refer to frequencies under some statistical model.

- p-value: what proportion of hypothetical samples from the *null hypothesis model*, would have a statistic at least as extreme as ours?
- Alpha: probability of rejecting the null hypothesis for data sampled from the *null hypothesis model*.
- Power: probability of rejecting the null hypothesis for data sampled from some *alternative model*.
- Sampling distribution: the probability distribution of a statistic given that it is sampled from *some model*.
- Confidence interval probability: probability that a confidence interval computed in this manner using samples from *some model* will contain the model parameter value.

Probabilities in null hypothesis significance testing refer to peculiar conditional probabilities:

- p-value:
  P(X › x.sample | null is true) P(X › x.sample | X~null)
- Alpha:
  P(significant | null is true)
- Power:
  P(significant | null is false)

- *Really important:*
  - *These do not give us the probability that the null is false: P(null is false | significant) !!*

Outline

- There is only one test.
- Central limit theorem, normal distribution, and sampling distribution of the sample mean
- Z-test
- Confidence intervals and NHST analogue
- Null hypothesis significance testing
- Power
- Wrap up