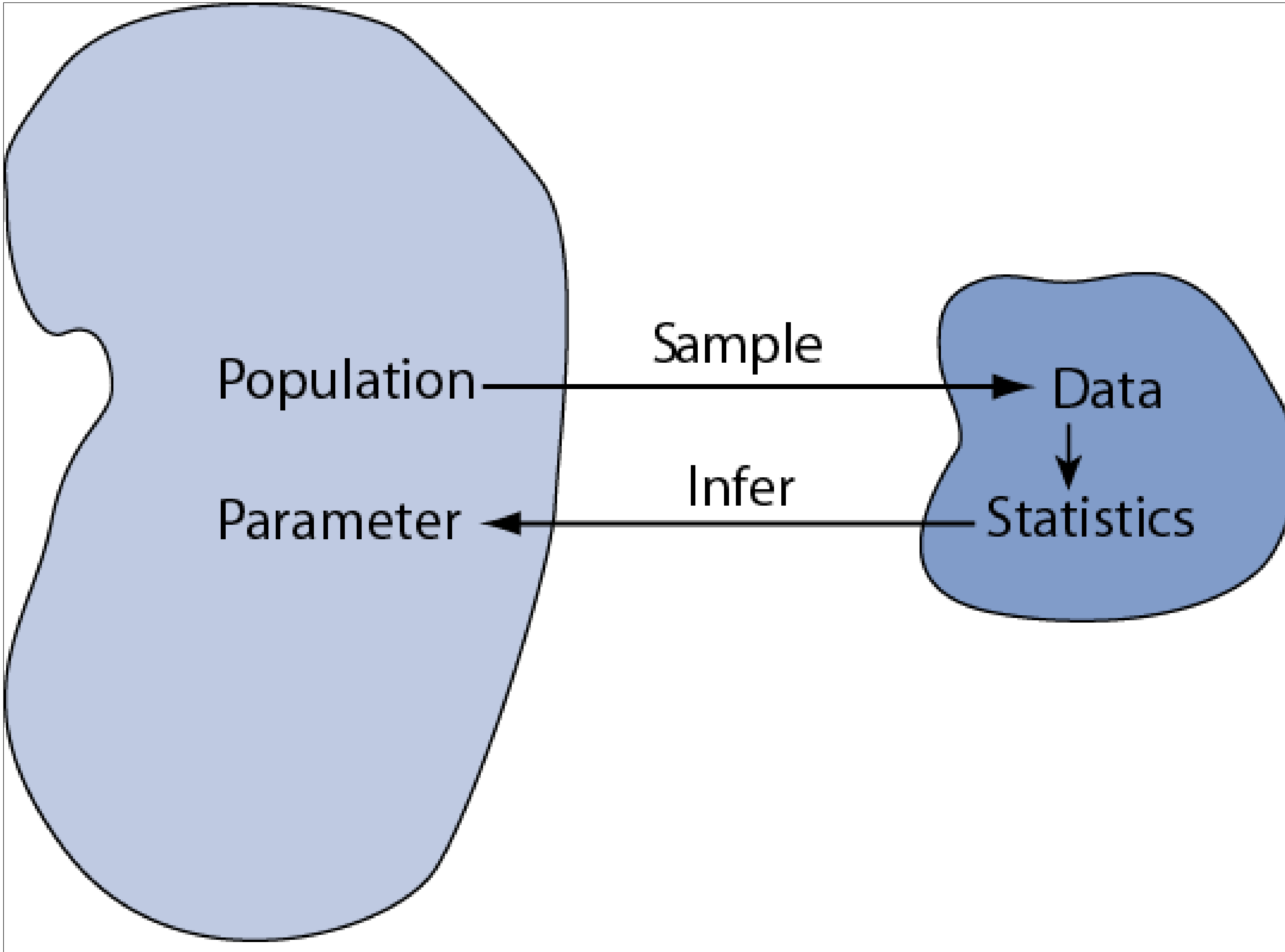
Introduction to statistics

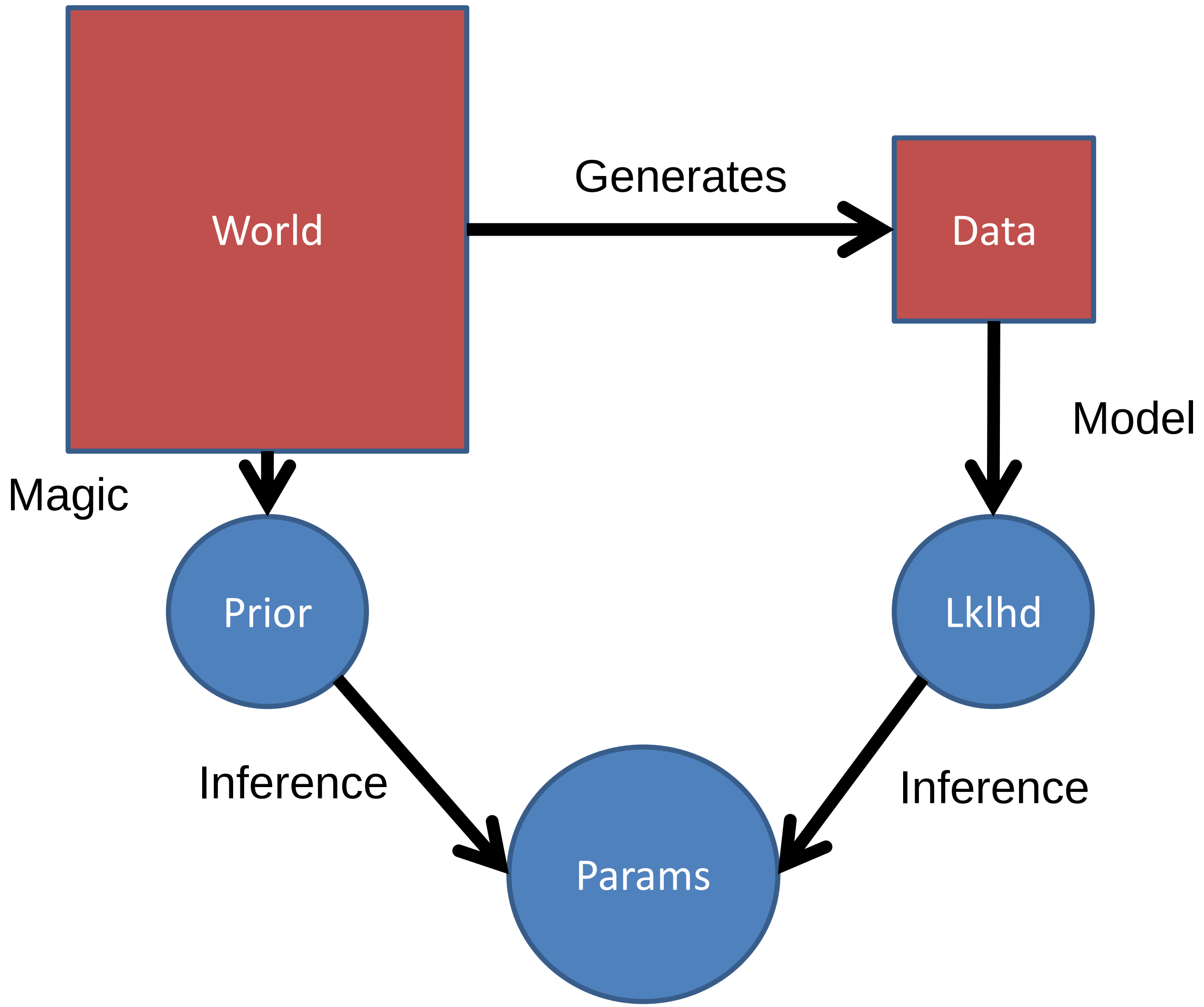
Nisheeth

Distinctions Between Parameters and Statistics

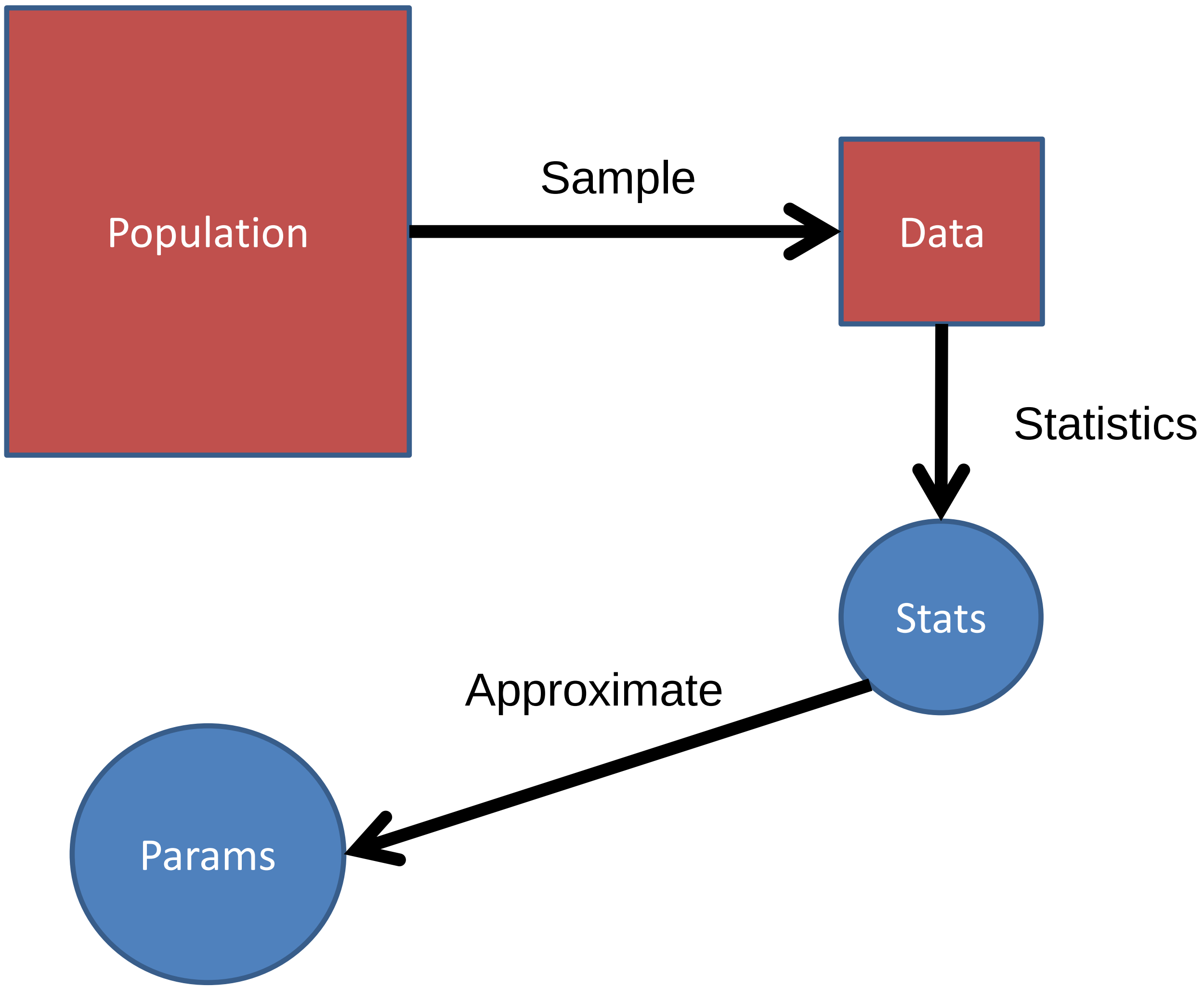
	Parameters	Statistics
Source	Population	Sample
Notation	Greek (e.g., μ)	Roman (e.g., *xbar*)
Vary	No	Yes
Calculated	No	Yes

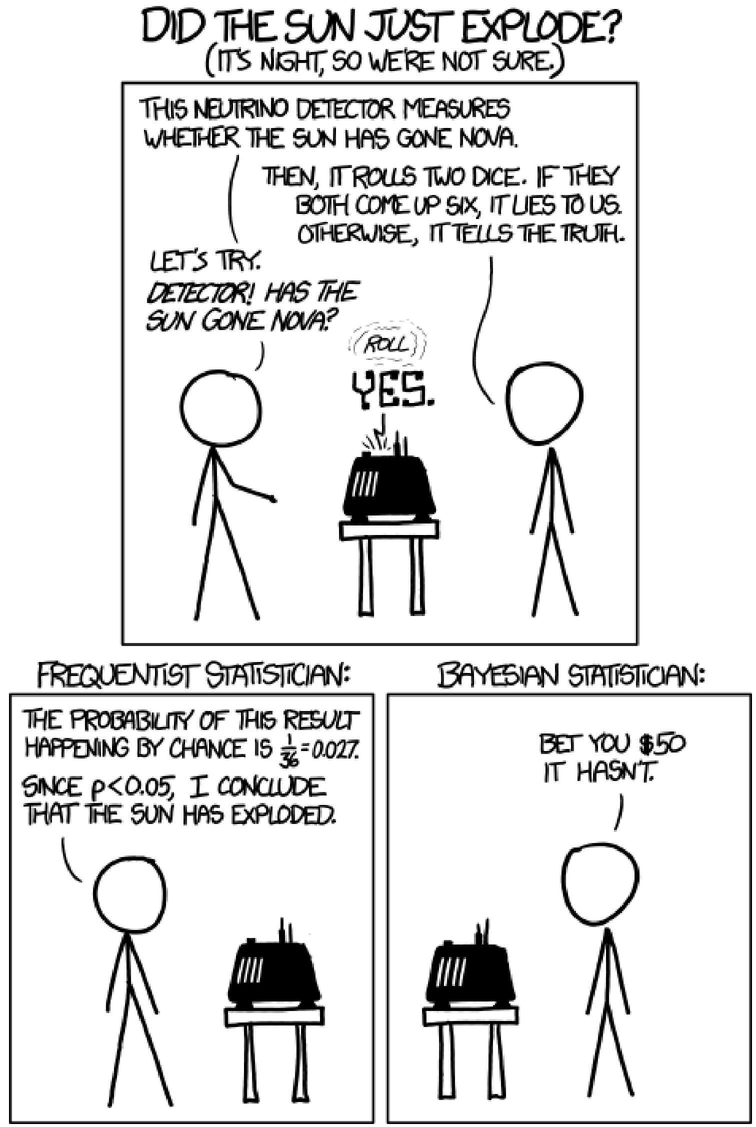

Sample
Population
Data
Infer
Parameter
Statistics

The Bayesian's universe



The statistician's universe





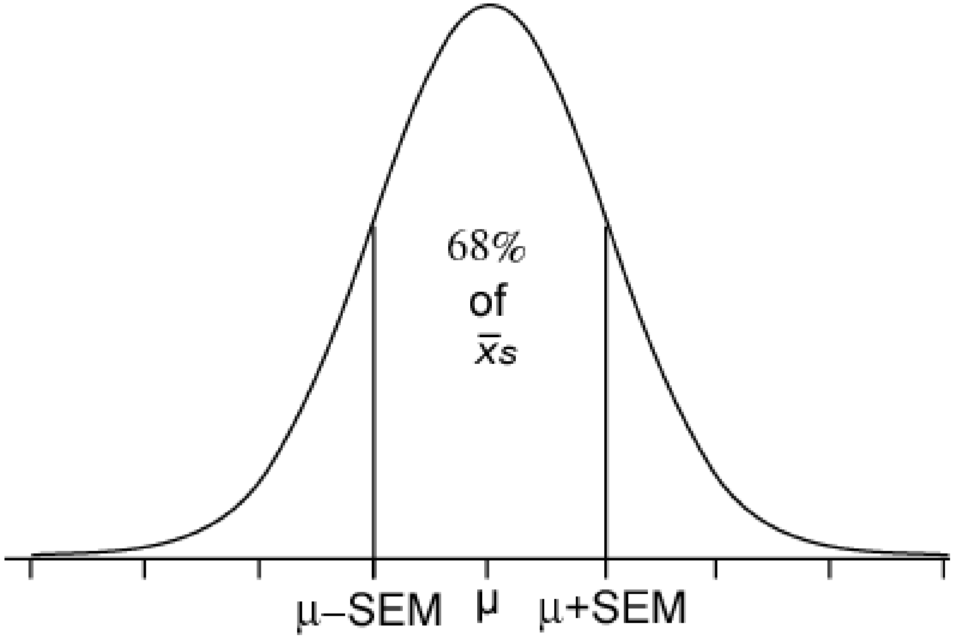
“The Reverend Bayes published posthumously. The field of statistics would greatly improve if all Bayesians were to follow his example”

Sampling Distributions of a Mean

The **sampling distributions of a mean (SDM)** describes the behavior of a sampling mean

$$\bar{x} \sim N\left(\mu, SE_{\bar{x}}\right)$$

$$\text{where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Hypothesis Testing

- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample
- The technique is introduced by considering a one-sample z test
- The procedure is broken into four steps
- *Each* element of the procedure must be understood

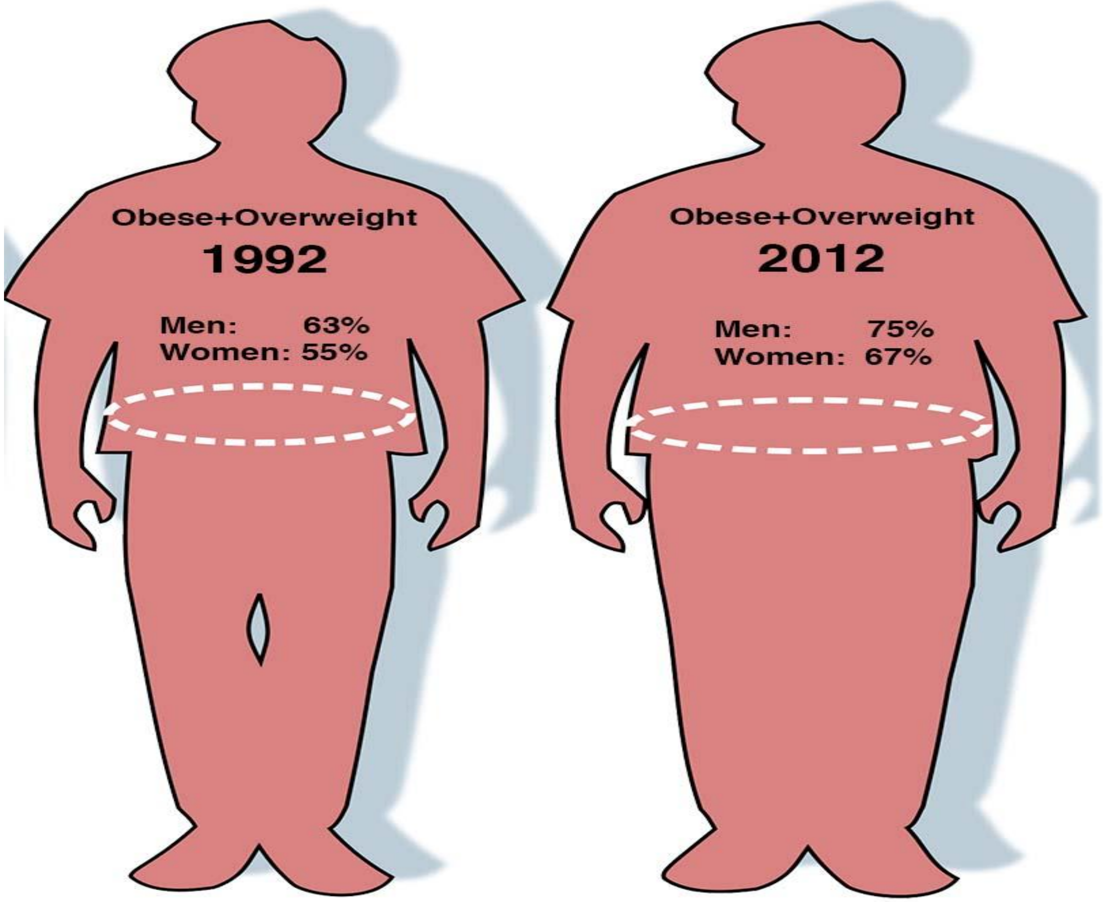
Hypothesis Testing Steps

A. Null and alternative hypotheses
B. Test statistic
C. P-value and interpretation
D. Significance level (optional)

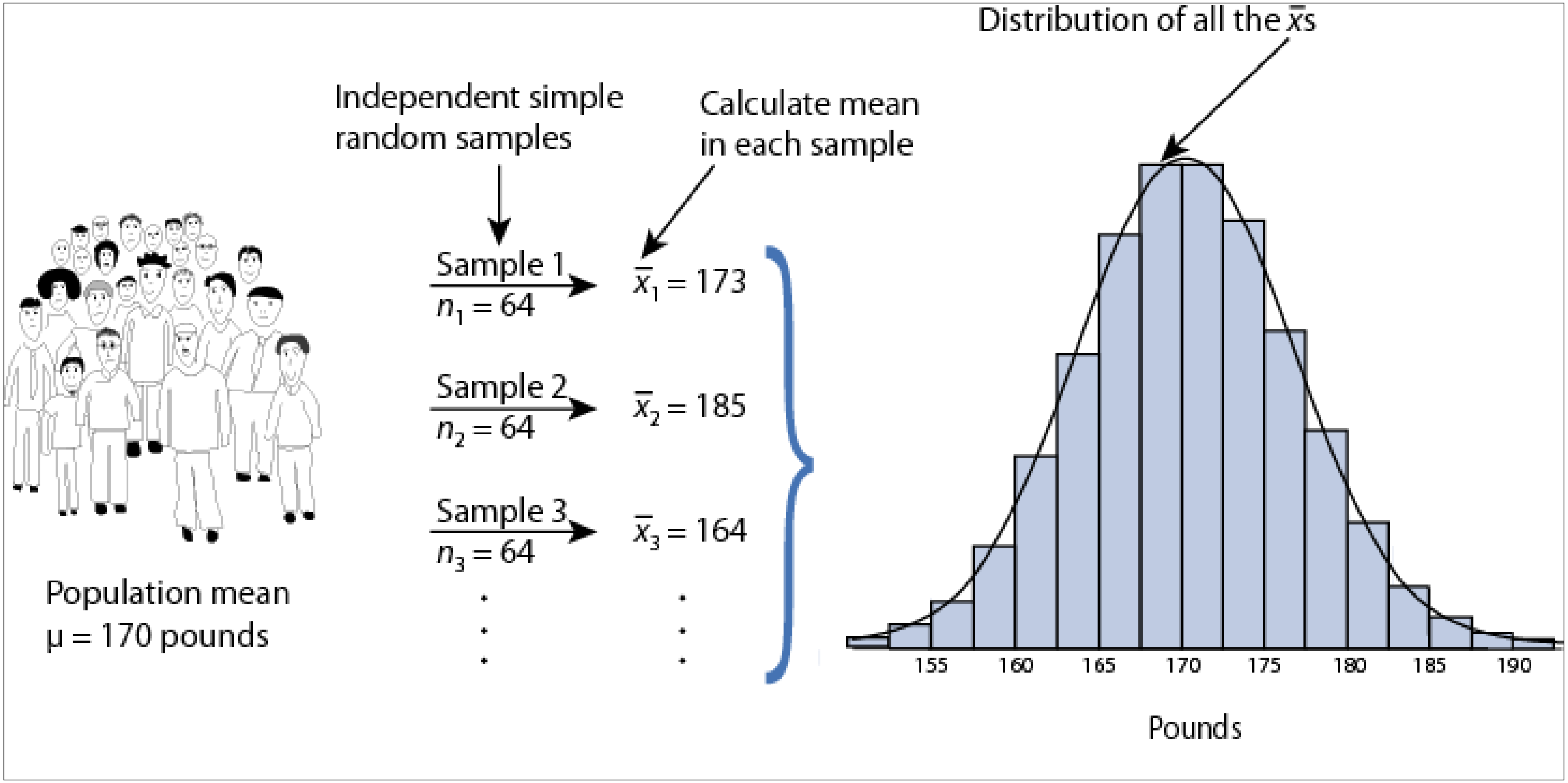
Null and Alternative Hypotheses

- Convert the research question to null and alternative hypotheses
- The **null hypothesis (H_0)** is a claim of "no difference in the population"
- The **alternative hypothesis (H_a)** claims "H_0 is false"
- Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction)

Illustrative Example: “Body Weight”



Reasoning



Sampling distribution of xbar under H_0: μ = 170 for n = 64 $\Rightarrow$ $\bar{x} \sim N(170,5)$

Test Statistic

This is an example of a one-sample test of a mean when σ is known. Use this statistic to test the problem:

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

$$\text{where } \mu_0 \equiv \text{population mean assuming } H_0 \text{ is true}$$

$$\text{and } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Illustrative Example: *z* statistic

- For the illustrative example, $\mu_0 = 170$
- We know $\sigma = 40$
- Take a random sample of *n* = 64. Therefore

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = 5$$

- If we found a sample mean of 173, then

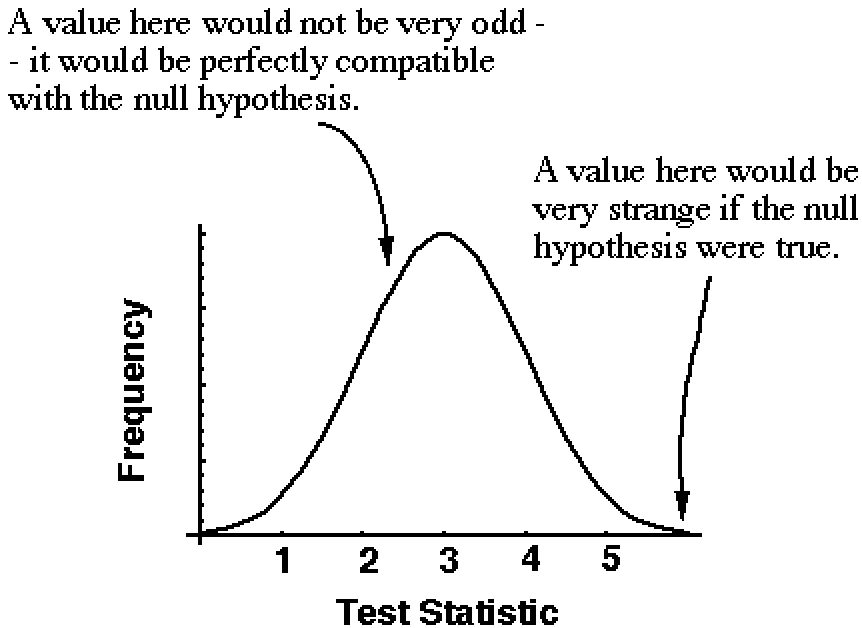
$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{173 - 170}{5} = 0.60$$

Illustrative Example: *z* statistic

If we found a sample mean of 185, then

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{185 - 170}{5} = 3.00$$

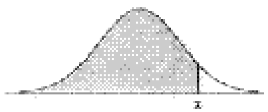
What is a test statistic?



Z-statistic $N(x|\mu,\sigma^2) \rightarrow N(z|0,1)$

$$p\left(\bar{x} \le \tilde{x} | N(x|\mu,\sigma^2)\right) = p(z \le \tilde{z} | N(z|0,1)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\tilde{z}} \exp\left(-\frac{z^2}{2}\right) dz$$

Tables of the Normal Distribution



Probability Content from -oo to Z

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

p-value

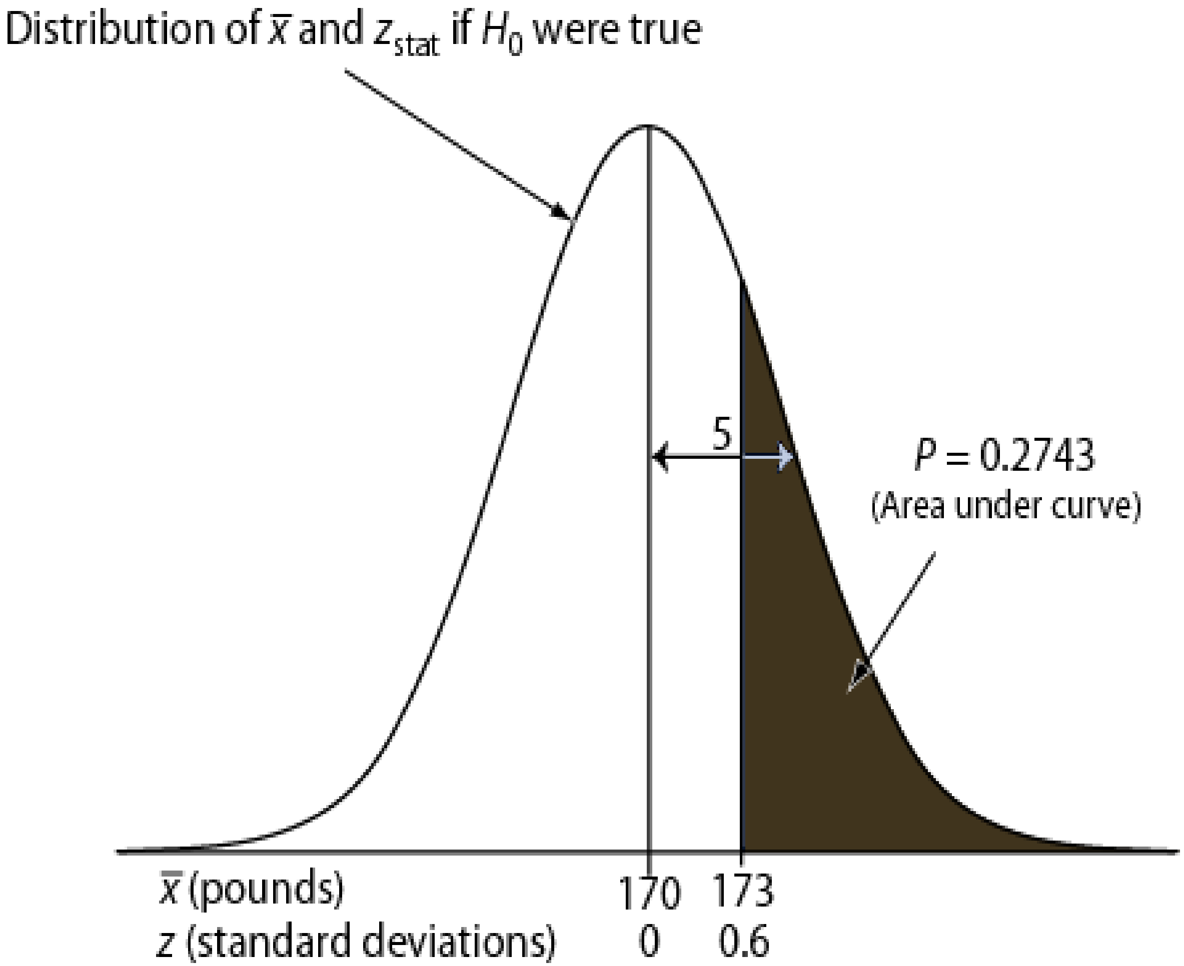
- The P-value answer the question: What is the probability of the observed test statistic or one more extreme **when H_0 *is true*?**
- This corresponds to the AUC in the tail of the Standard Normal distribution beyond the $z_{stat.}$
- Convert z statistics to P-value :

  For H_a: $\mu > \mu_0 \Rightarrow P = \Pr(Z > z_{stat})$ = right-tail beyond z_{stat}

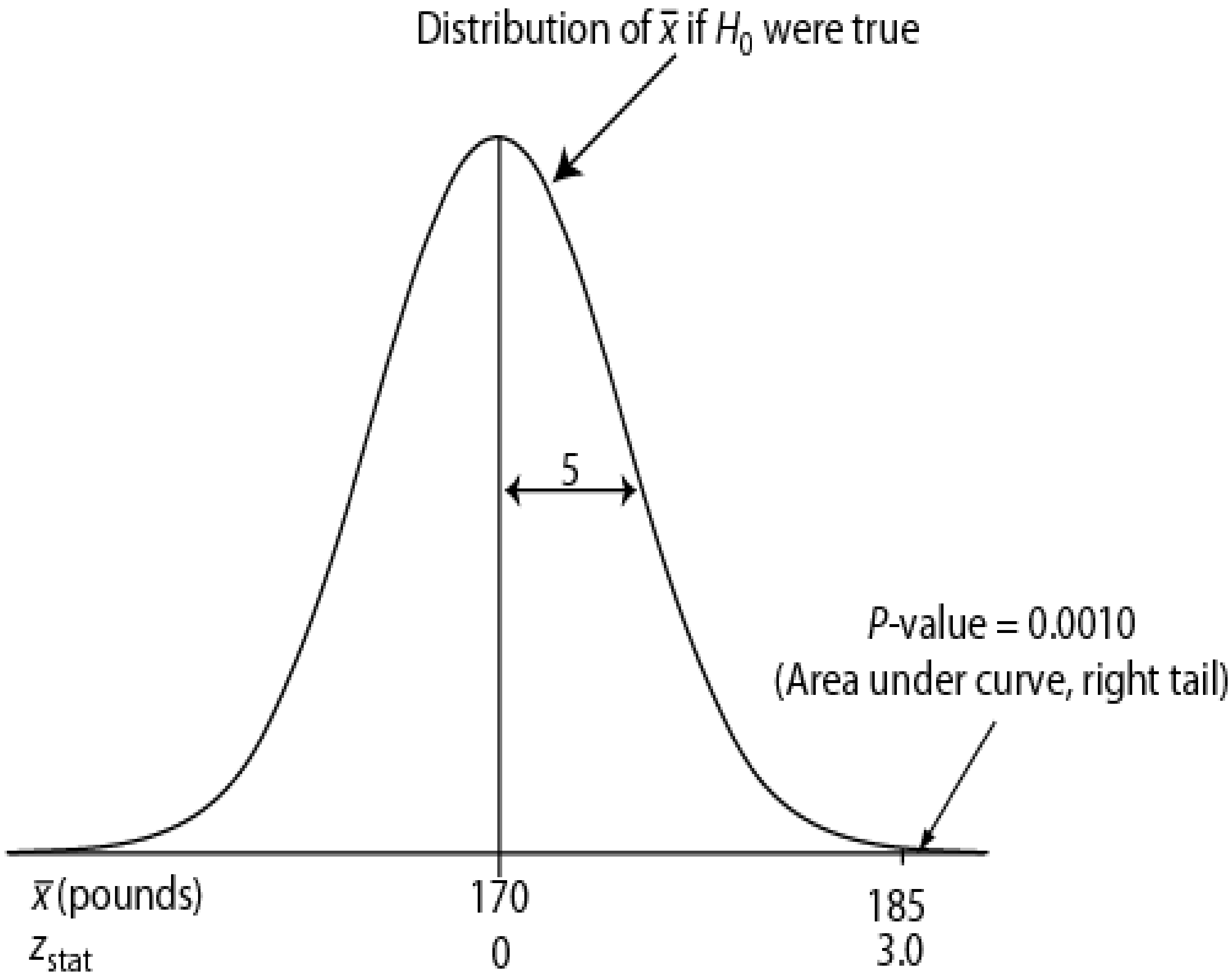
  For H_a: $\mu < \mu_0 \Rightarrow P = \Pr(Z < z_{stat})$ = left tail beyond z_{stat}

  For H_a: $\mu \neq \mu_0 \Rightarrow P = 2 \times$ one-tailed P-value

One-sided P-value for z_{stat} of 0.6

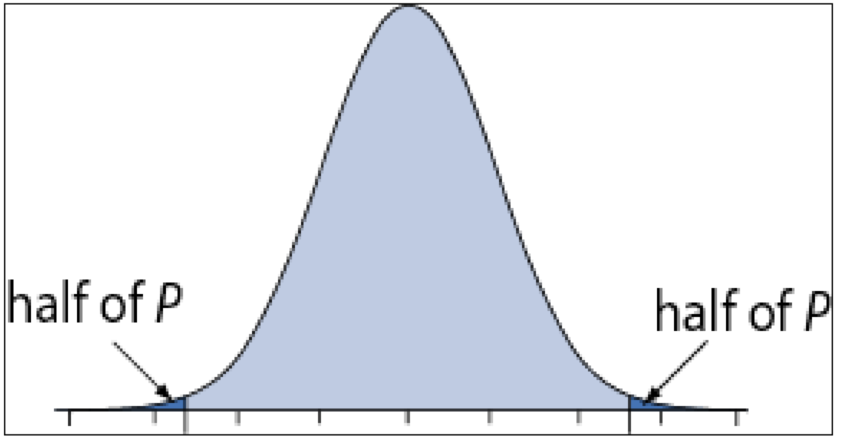


One-sided P-value for z_{stat} of 3.0



Two-Sided *P*-Value

- One-sided $H_a \Rightarrow$ AUC in tail beyond z_{stat}
- Two-sided $H_a \Rightarrow$ consider potential deviations in both directions $\Rightarrow$ double the one-sided *P*-value



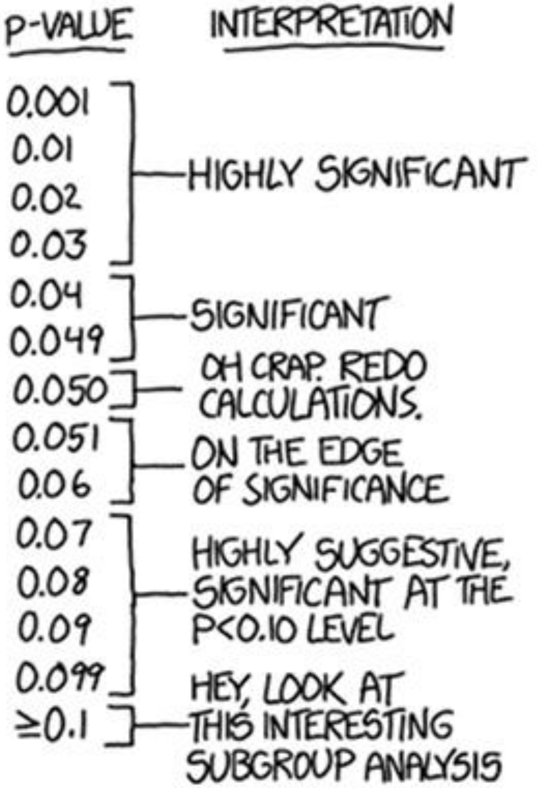
Examples: If one-sided *P* = 0.0010, then two-sided *P* = 2 × 0.0010 = 0.0020. If one-sided *P* = 0.2743, then two-sided *P* = 2 × 0.2743 = 0.5486.

Interpretation

- *P*-value answer the question: What is the probability of the observed test statistic … **when H_0 *is true*?**
- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against H_0
- Small *P*-value $\Rightarrow$ strong evidence

α-Level (Used in some situations)

- Let $\alpha \equiv$ probability of erroneously rejecting H_0
- Set α threshold (e.g., let $\alpha = .10$, .05, *or whatever*)
- Reject H_0 when $P \leq \alpha$
- Retain H_0 when $P > \alpha$
- Example: Set $\alpha = .10$. Find $P = 0.27 \Rightarrow$ retain H_0
- Example: Set $\alpha = .01$. Find $P = .001 \Rightarrow$ reject H_0



(Summary) One-Sample *z* Test

A. Hypothesis statements

H_0: $\mu = \mu_0$ vs.

H_a: $\mu \neq \mu_0$ (two-sided) or

H_a: $\mu < \mu_0$ (left-sided) or

H_a: $\mu > \mu_0$ (right-sided)

B. Test statistic

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \text{ where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

C. P-value: convert *z* to p value

D. Significance statement depending on α

Another example: IQ testing

- Let X represent Weschler Adult Intelligence scores (WAIS)
- Typically, $X \sim N(100, 15)$
- Take i.i.d. samples of $n = 9$ from a population
- Data $\Rightarrow$ {116, 128, 125, 119, 89, 99, 105, 116, 118}
- Calculate sample mean = 112.8
- Does sample mean provide strong evidence that population mean $\mu > 100$?

Example

A. Hypotheses:

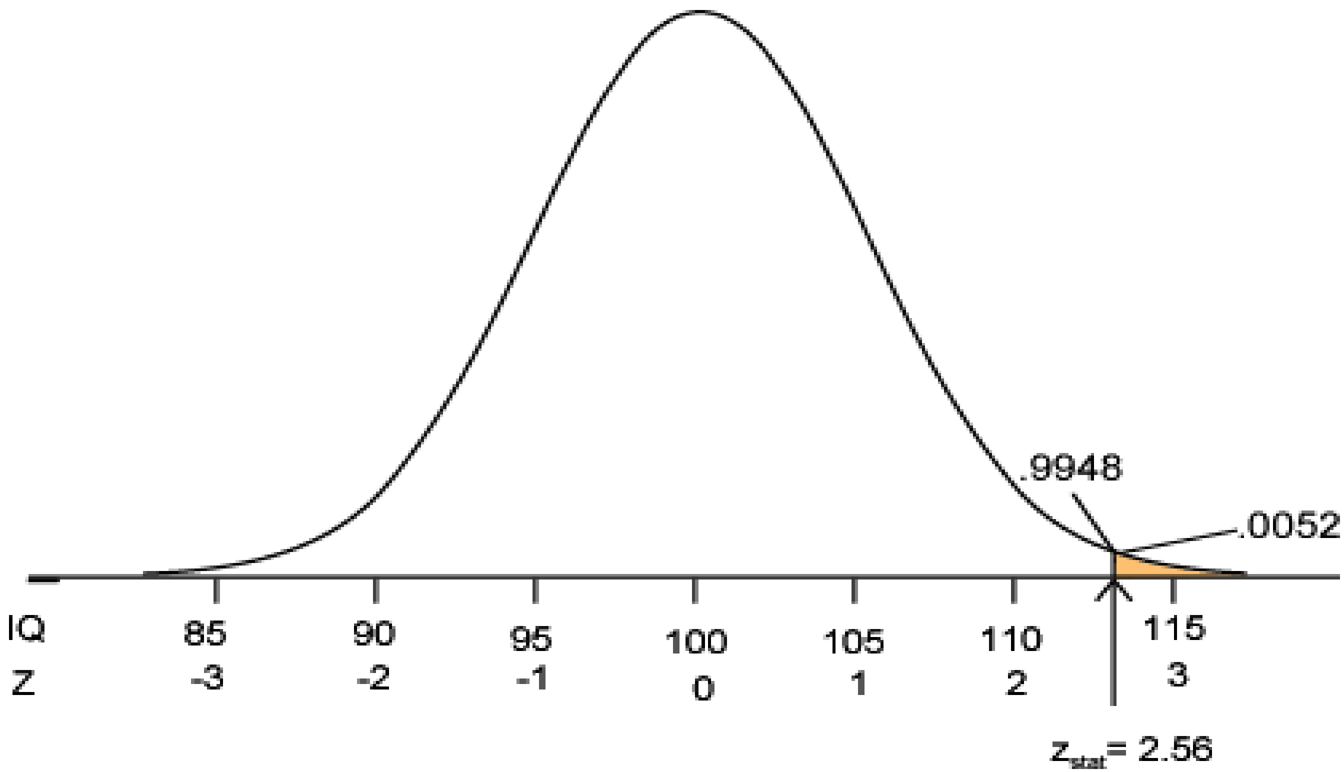
H_0: μ = 100 versus
H_a: μ > 100 (one-sided)
H_a: μ ≠ 100 (two-sided)

B. Test statistic:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{9}} = 5$$

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{112.8 - 100}{5} = 2.56$$

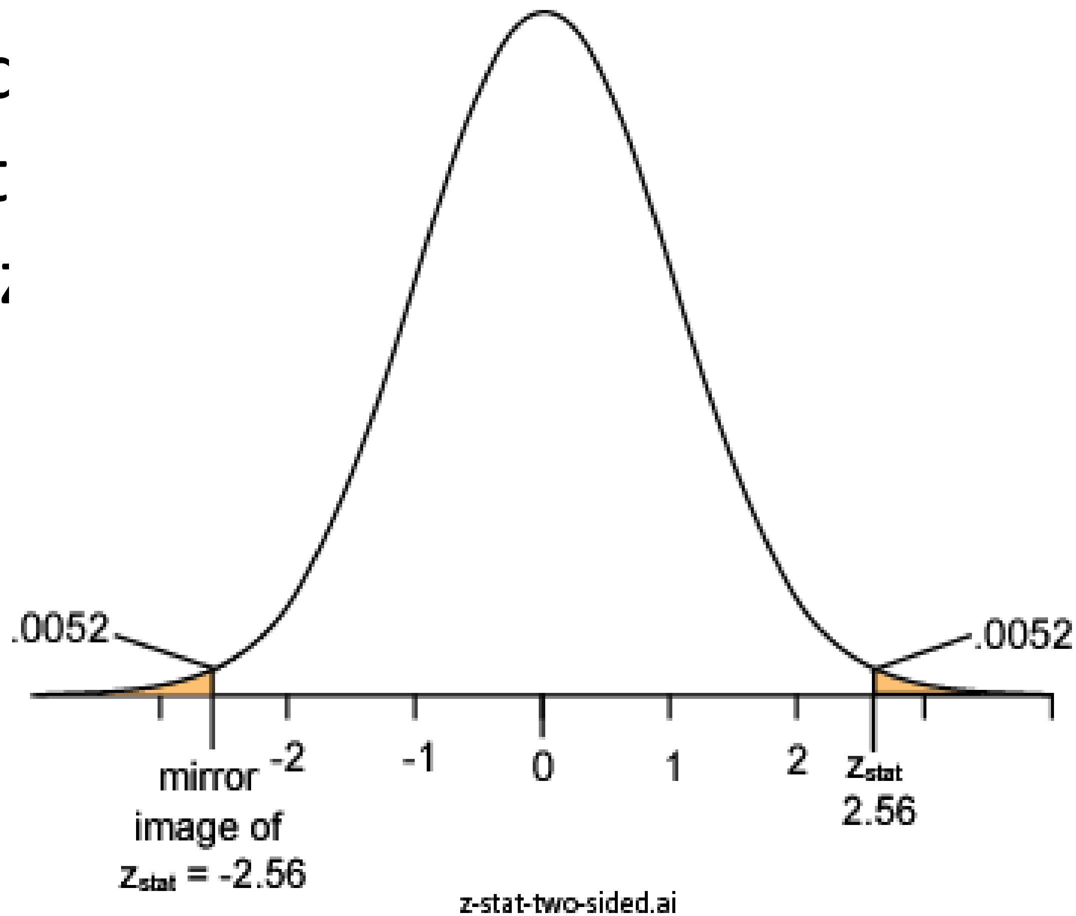
C. *P*-value: $P = \Pr(Z \geq 2.56) = 0.0052$



P =.0052 ⇒ it is unlikely the sample came from this null distribution ⇒ strong evidence against H_0

Two-sided *p*-value

- H_a: $\mu \neq 100$
- Considers random deviations "up" and "down" from $\mu_0 \Rightarrow$ t above and below $\pm z$
- Thus, two-sided *P* $= 2 \times 0.0052$ $= 0.0104$



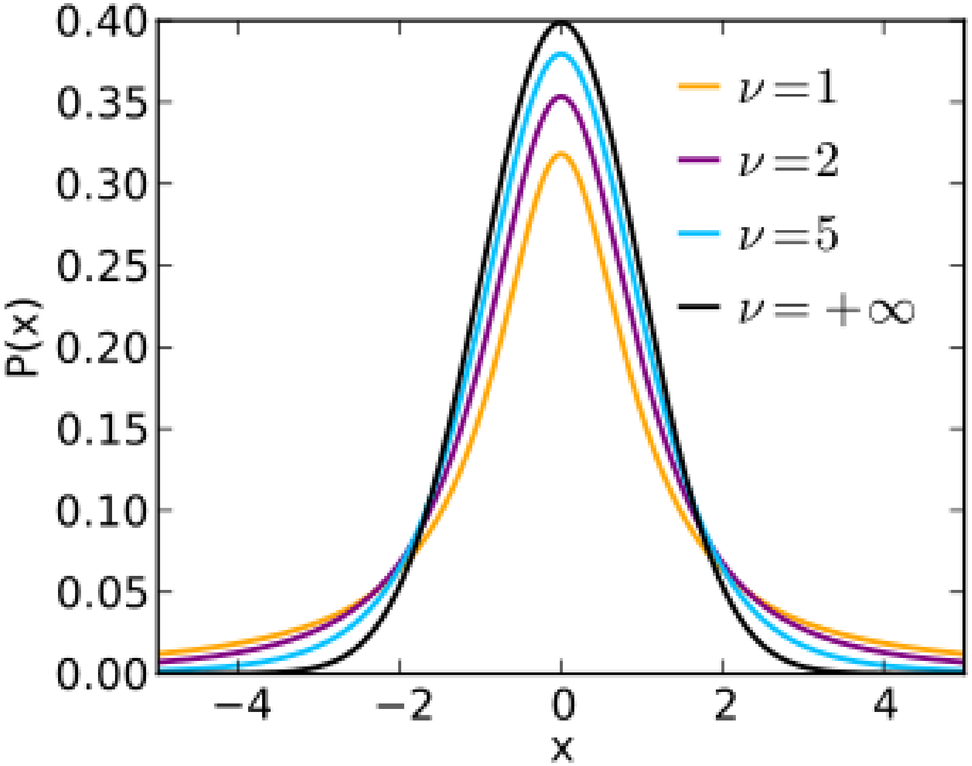
Conditions for z test

- σ known (not from data)
- Population approximately normal or large sample ($n>30$)
- Data i.i.d.

T tests

- The z test conditions seldom hold in practice
  - We don't often know the population variance
  - The sample size can be small
- We use a T test instead
  - Assumes the sampling distribution is a t-distribution
  - $x \sim \frac{\Gamma(\frac{\upsilon+1}{2})}{\sqrt{\upsilon\pi}\,\Gamma(\frac{\upsilon}{2})}\left(1+\frac{x^2}{\upsilon}\right)^2$

T distribution



$$T_{n-1}(\bar{x}) = \frac{\bar{x} - \mu}{S/\sqrt{n}}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Can you say why?

(Summary) T Test

A. Hypothesis statements

H_0: $\mu = \mu_0$ vs.

H_a: $\mu \neq \mu_0$ (two-sided) or

H_a: $\mu < \mu_0$ (left-sided) or

H_a: $\mu > \mu_0$ (right-sided)

B. Calculate Test statistic

$$T_{n-1}(\bar{x}) = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \text{where,} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

C. P-value: convert T to p value

D. Significance statement depending on α

Test statistics

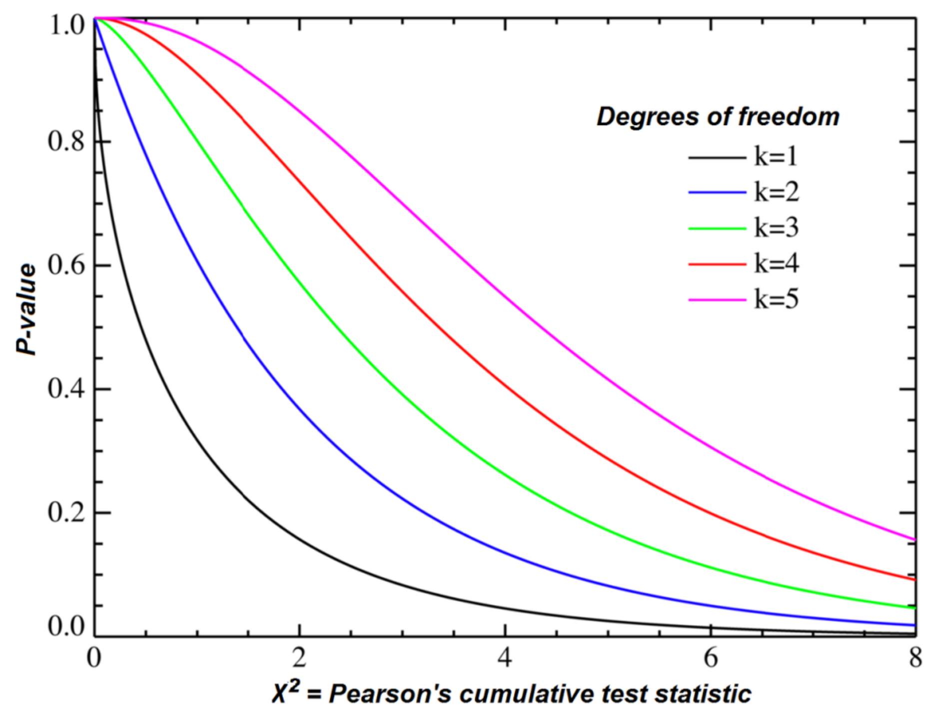
Type Of Test	Purpose	Example	Equation	Comment
Z Test	Test if the average of a single population is equal to a target value	Do babies born at this hospital weigh more than the city average	$Z = \frac{\bar{x} - u_0}{\frac{\sigma}{\sqrt{n}}}$	Z test does not need df σ = population standard deviation
1 Sample T-Test	Test if the average of a single population is equal to a target value	Is the average height of male college students greater than 6.0 feet?	$t = \frac{\bar{x} - u_0}{\frac{s}{\sqrt{n}}}$ $df = n - 1$	s = sample standard deviation
Paired T-Test	Test if the average of the differences between paired or dependent samples is equal to a target value	Weigh a set of people. Put them on a diet plan. Weigh them after. Is the average weight loss significant enough to conclude the diet works?	$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$ $df = n - 1$	d bar = average difference between samples s = sample deviation of the difference n = count of one set of the pairs (don't double count)
2 Sample T-Test Pooled Variance	Test if the difference between the averages of two independent populations is equal to a target value	Do cats eat more of type A food than type B food	$df = n_1 + n_2 - 2$ $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	n1, n2 = count of sample 1, 2
2 Sample T-Test Unpooled Variance	Test if the difference between the averages of two independent populations is equal to a target value	Is the average speed of cyclists during rush hour greater than the average speed of drivers	$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

Test statistic construction

- Proportions as random variables
- What does the sampling distribution of the mean look like?
- What will the test statistic look like?

Different set of tests for variance

- Sampling distribution for variance looks very different
- Hypothesis testing logic stays the same
  - – Hypothesize
  - – Calculate
  - – Interpret



(Summary) Statistical hypothesis testing

A. Hypothesis statements
   H_0: $\mu = \mu_0$ vs.
   H_a: $\mu \neq \mu_0$ (two-sided) or
   H_a: $\mu < \mu_0$ (left-sided) or
   H_a: $\mu > \mu_0$ (right-sided)
B. Calculate Test statistic
C. P-value: convert *test statistic* to p value
D. Significance statement depending on α

PRACTICAL CONSIDERATIONS

Multiple comparisons

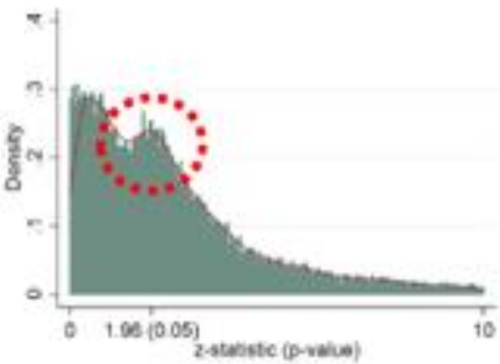
- Let’s say we’re doing IQ testing department-wise
  - 7 departments
  - $\alpha = 0.05$
  - Run $C(7,2) = 21$ pairwise T-tests on the data
  - CSE > EE
- How to interpret?
- Bonferroni correction for FWER $\alpha^* = \alpha/\#comparisons$

P-hacking

Economics

Brodeur et al (AEJ:A, in press)

"Star Wars: The empirics strike back"

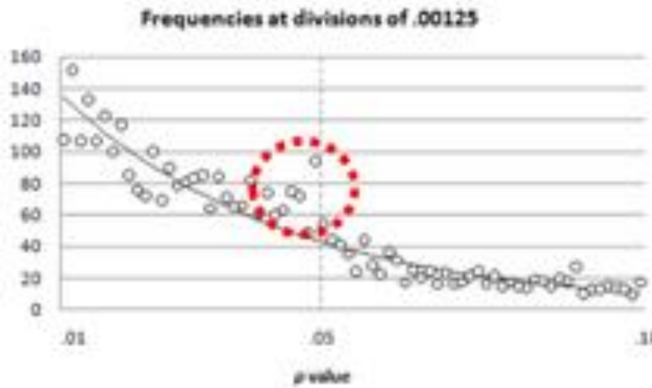


(b) De-rounded distribution of z-statistics.

Psychology

Masicampo Lalande (QJEP, 2012)

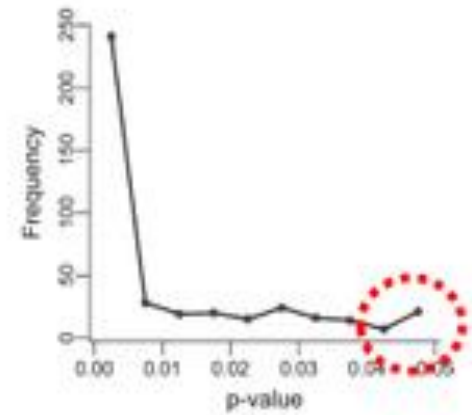
"A peculiar prevalence of p values just below .05"



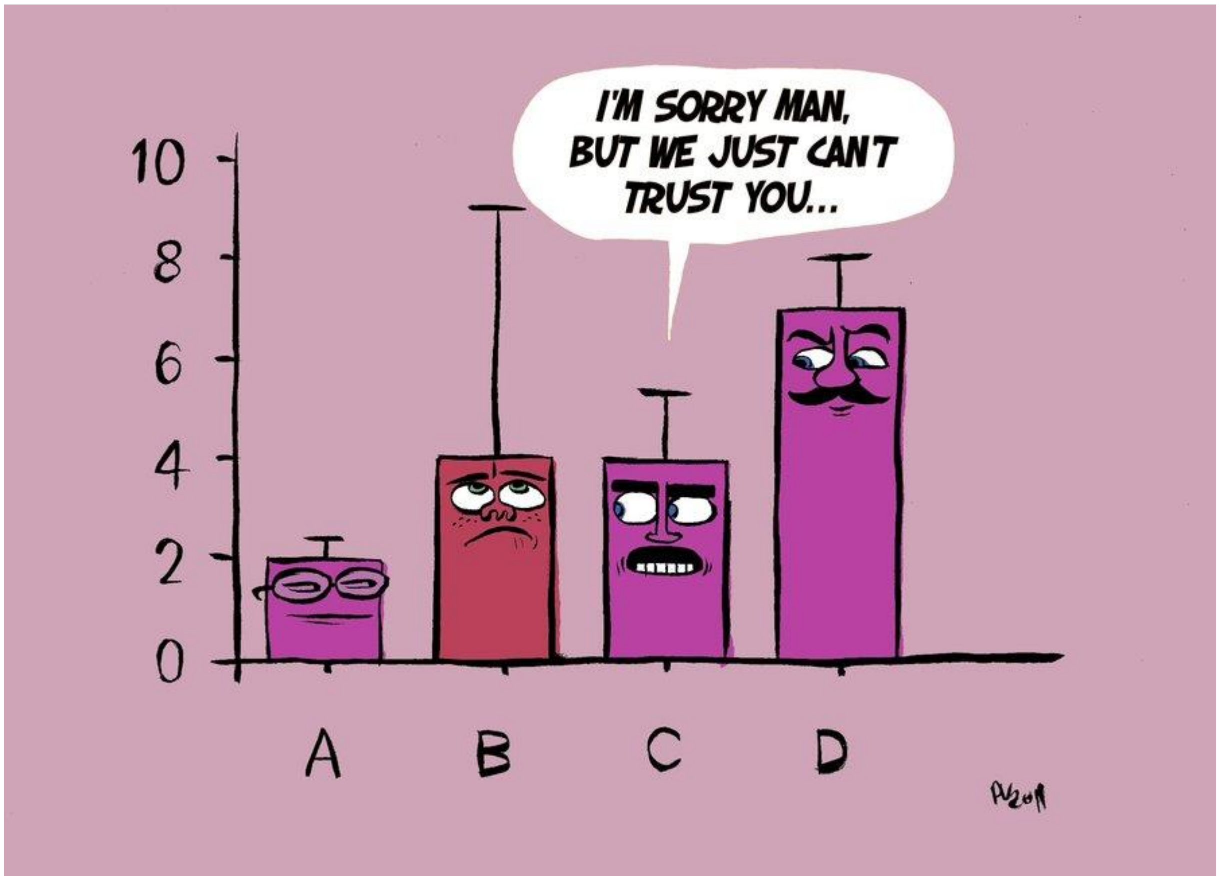
Biology

Head et al (PLOS Biology 2015)

"Extent and Consequences of P-Hacking in Science"



Reading error bars



Reading error bars

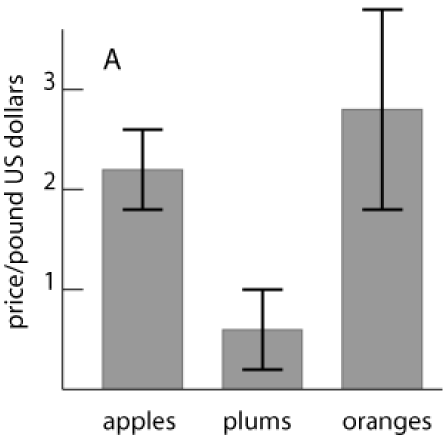


Fig 1. Inconsistency of prices for common fruit items among twelve major cities (bars represent standard deviation).

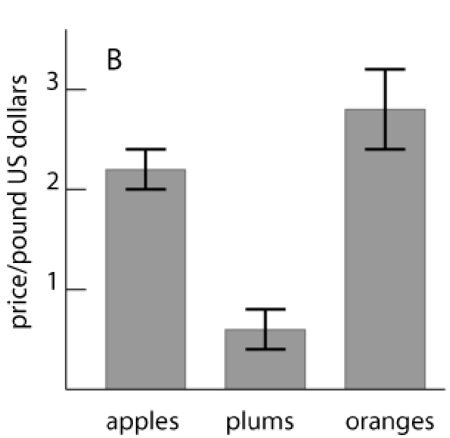


Fig 1. Cost comparison among three common fruit items based on prices in twelve major cities (bars represent standard deviation of the mean).

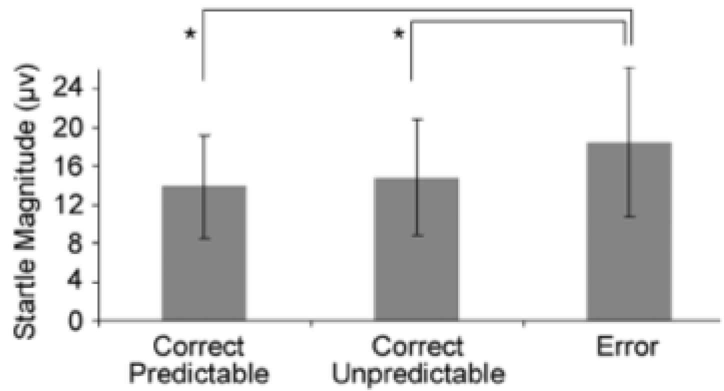


Fig. 1. Mean startle magnitude as a function of trial type. Error bars represent $\pm$ 2.04 standard errors, the 95% confidence interval. Asterisks indicate significant differences between trial types, $^*p < .01$.

Use SD when you want to describe the dataset

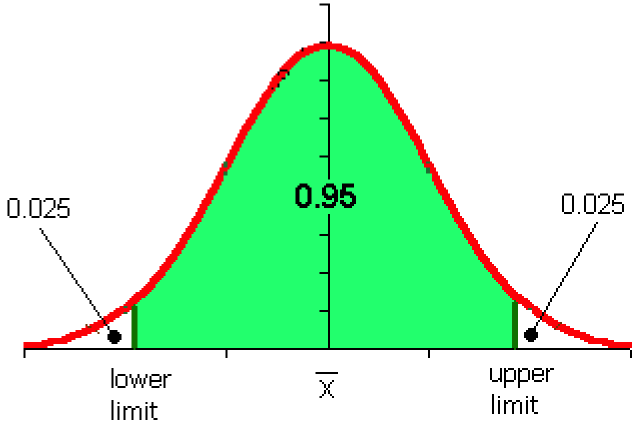
Use SE when you want to describe the precision of your estimate of the mean

Use CI when your SE is small enough for the CI to look impressive to the viewer!

Use of CI is strongly encouraged to report estimate precision. Be brave and use it

Confidence intervals

- Statistical estimates of population parameters are typically presented with *confidence intervals*
- $95\%\ CI = Z(p = 0.05) \times SE$
- Half the CI is sometimes called the *margin of error*
- How to set up data collection for a known margin of error?



Confidence level	Z value
90%	1.65
95%	1.96
99%	2.58
99,9%	3.291

Sample size calculation: simple

- Simplest case: one sample Z test
- Assume we want to find the population mean with margin of error at most W
- p = 0.05
- Then we want a CI of 2W and an SE of about W
- $W = \frac{\sigma}{\sqrt{n}}$, solve for n
- $n = \frac{\sigma^2}{W^2}$

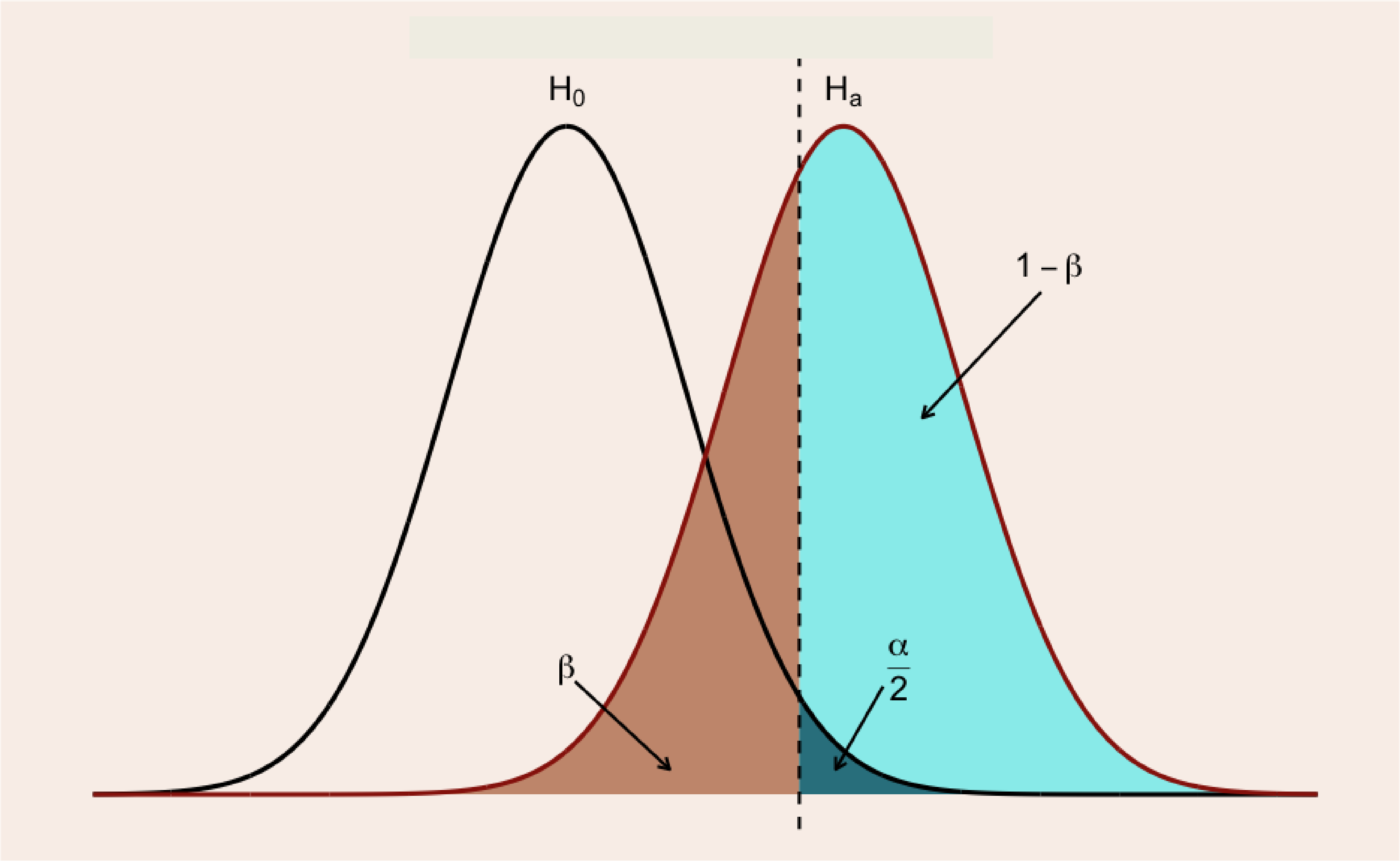
Cool trick: bootstrapping

- Power calculations depend on knowing the population standard deviation
  - Fine for Z tests, but what about the others?
- We use bootstrapping to estimate the population standard deviation from the data
  - Resample the data by drawing a dataset from the existing dataset randomly with replacement
  - Compute statistics
  - Repeat 1000 times
  - You get an empirical distribution on the statistic
  - Directly use the standard deviation of this distribution as the standard error of the statistic

Is this enough?

- Set up competing hypotheses
- Specify significance level
- Calculate confidence bound for test statistic
  - Use bootstrap if population variance is unknown
- Calculate effective sample size
- Collect data
- Calculate test statistic
- Output result

Not enough



Power

Two types of decision errors:

Type I error (FN) = erroneous rejection of true H_0

Type II error (FP) = erroneous retention of false H_0

	Truth	
Decision	H_0 true	H_0 false
Retain H_0	Correct retention	Type II error
Reject H_0	Type I error	Correct rejection

α ≡ probability of a Type I error

β ≡ Probability of a Type II error

Power

- $\beta \equiv$ probability of a Type II error

  β = Pr(retain H_0 | H_0 false)

- $1 - \beta$ = "Power" $\equiv$ probability of avoiding a Type II error

  $1 - \beta$ = Pr(reject H_0 | H_0 false)

Power calculation

- Calculate probability of a Type II error
  - Pr(null is not rejected|null is false)
  - $p(z < z_{crit} | z = \frac{x - \mu_1}{\sigma/\sqrt{n}}, x \sim N(\mu_1, \sigma))$
  - Power is 1 – p(Type II error)
- Calculate $x_{crit} = \mu_0 + z_{crit}\, \sigma/\sqrt{n}$
  - Power = $p(x > x_{crit} | x \sim N(\mu_1, \sigma))$
  - Power = $p\left(z \geq \frac{x_{crit} - \mu_1}{\sigma/\sqrt{n}}\right) = 1 - \Phi(\frac{x_{crit} - \mu_1}{\sigma/\sqrt{n}})$

Calculating Power: Example

A study of n = 16 retains H_0: μ = 170 at α = 0.05 (two-sided); σ is 40. What is the power of test's conditions to identify a population mean of 190?

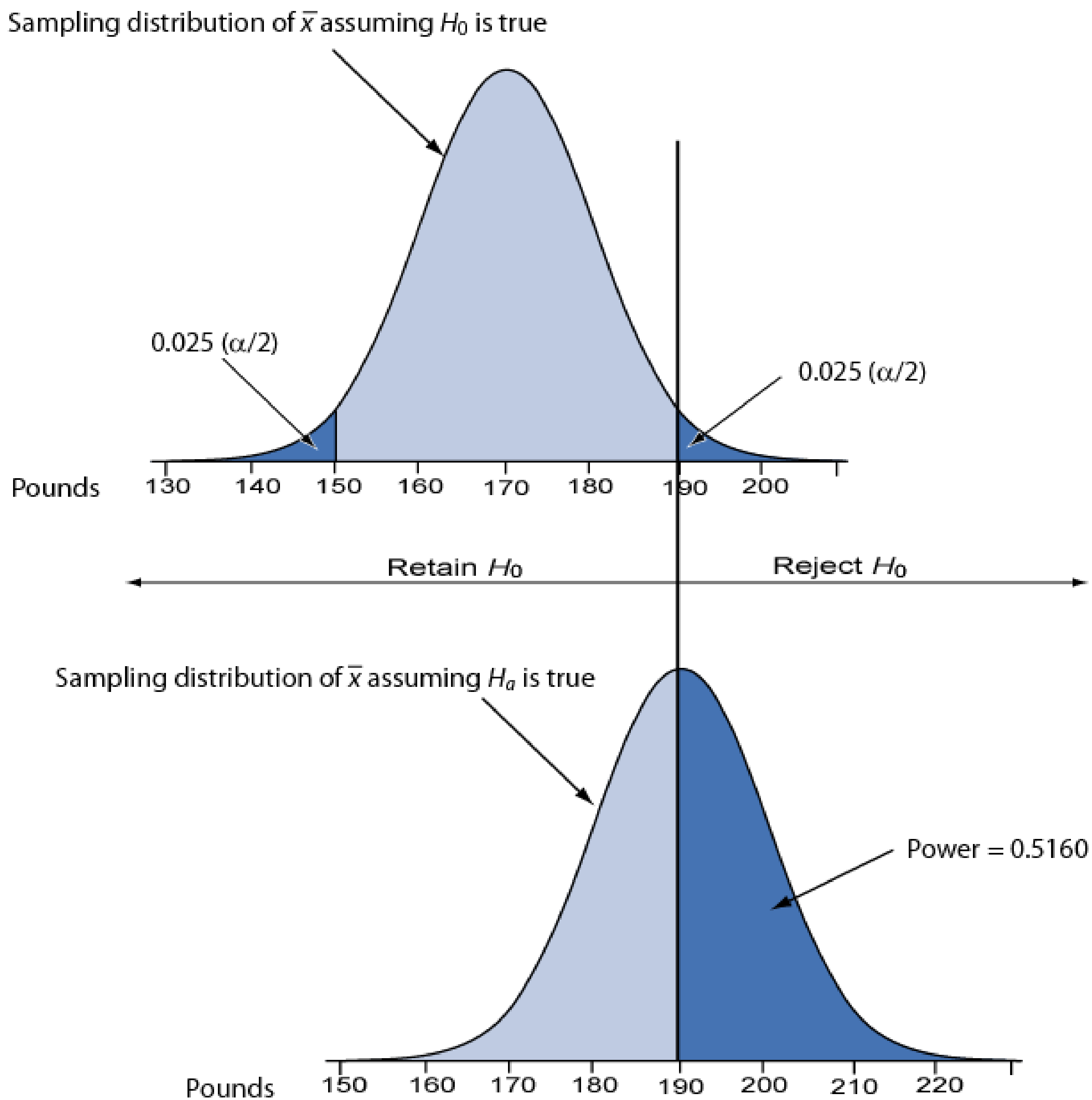
$$x_{crit} = 170 + 1.96 \times \frac{40}{\sqrt{16}} = 189.6$$

$$Power = 1 - \Phi(-0.04) = 0.5160$$

Can you find the power if we used 100 samples instead?

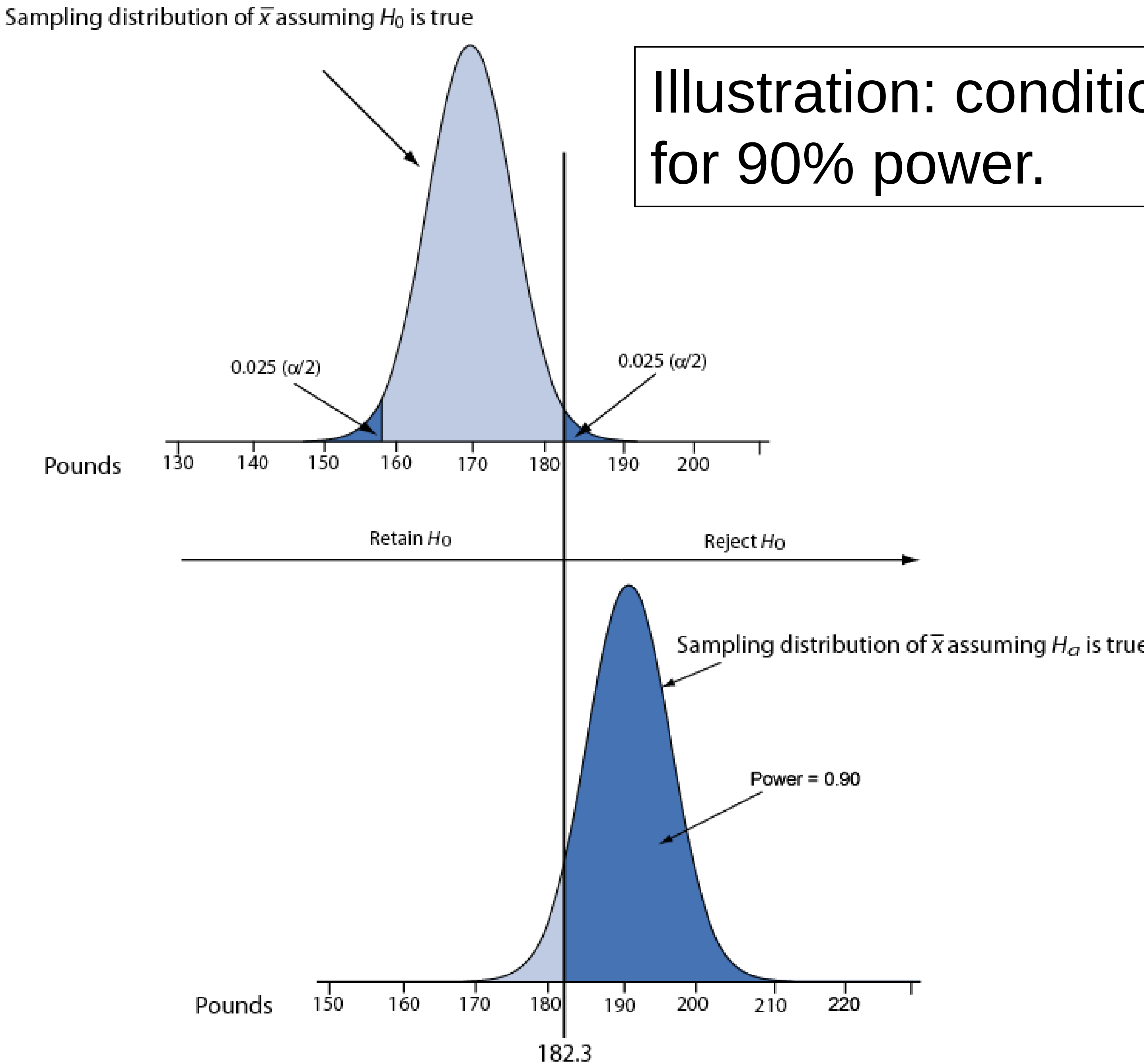
$$x_{crit} = 170 + 1.96 \times \frac{40}{\sqrt{100}} = 177.68$$

$$Power = 1 - \Phi(-3.08) = 0.999$$



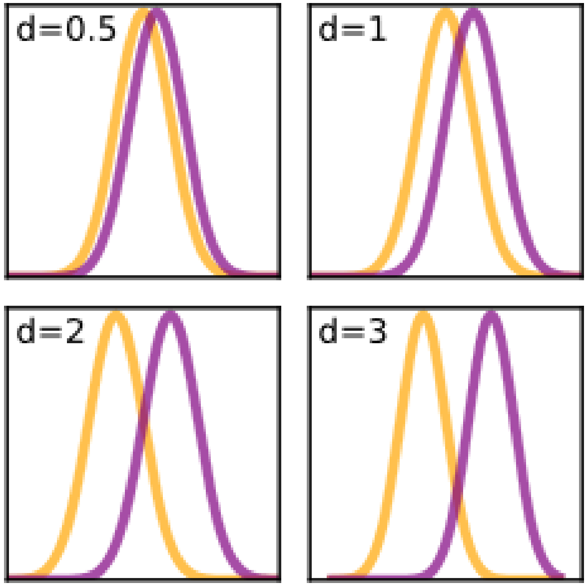
Sampling distribution of $\bar{x}$ assuming H_0 is true
0.025 ($\alpha/2$)
0.025 ($\alpha/2$)
Pounds
130
140
150
160
170
180
190
200
Retain H_0
Reject H_0
Sampling distribution of $\bar{x}$ assuming H_a is true
Power = 0.5160
Pounds
150
160
170
180
190
200
210
220

Illustration: conditions for 90% power.



Effect size

- Frequently measured using Cohen's d
- $d = \frac{\mu_1 - \mu_0}{s}$
- Bigger effect sizes easier to discriminate
- Estimate via pilots

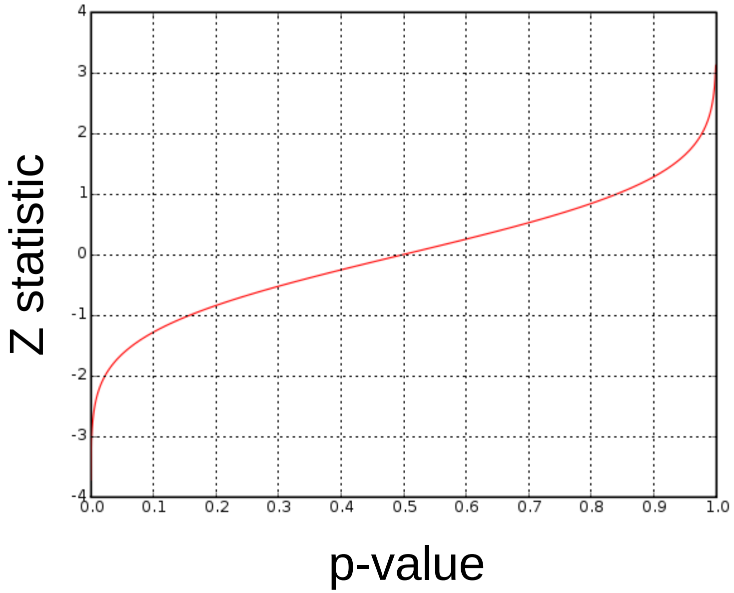


Summary of statistical calculations

- Given any 3 of α, β, d and n, we can calculate the fourth
  - Can calculate expected margin of error a test
  - Can calculate expected power of a test
  - Can calculate minimum discriminable effect size of a test
  - Can calculate required sample size of a test

Sample size calculation: complex

- We know α, β, d
- Calculate n
- Need to know values for the normal quantile function
- $n = \frac{\left(\Phi^{-1}(\beta) - z_{crit}\right)^2}{d^2}$
- Very important
  - Science experiments
  - Costly data collection



Review: statistics

- The language of statistics
  - Describes a universe where we sample datasets from a population
- Interesting properties are proved for sampling distributions of parameter estimates
- Statistical hypothesis testing
  - Helps us decide if a sample belongs to a population
- A priori calculation of important statistical properties can help design better studies
  - Power, sample size, effect size