Statistical Thinking for Forensic Practitioners

Part 6: Inference (Video 1)

Dr. Alicia Carriquiry

csafe
Center for Statistics and Applications in Forensic Evidence

Motivation - ASTM 2927-16

- ASTM 2927-16: Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons
  - Introduction. "One objective of a forensic glass examination is to compare glass samples to determine if they may be discriminated using their physical, optical or chemical properties (for example, color, refractive index (RI), density, elemental composition)..... The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass."
  - The "Big Picture" applies in this situation as well
  - Now two populations (one corresponding to known source and one corresponding to questioned source)
  - Question of interest is whether these populations differ in important ways (are distinguishable)

Motivation - ASTM 2927-16

- 11. Calculation and Interpretation of Results

  11.1. The procedure below shall be followed to conduct a forensic glass comparison when using the recommended match criteria:

  11.1.1. For the Known source fragments, using a minimum of 9 measurements (from at least 3 fragments, if possible), calculate the mean for each element.

  11.1.2. Calculate the standard deviation for each element. This is the Measured SD.

  11.1.3. Calculate a value equal to at least 3% of the mean for each element. This is the Minimum SD.

  11.1.4. Calculate a match interval for each element with a lower limit equal to the mean minus 4 times the SD (Measured or Minimum, whichever is greater) and an upper limit equal to the mean plus 4 times the SD (Measured or Minimum, whichever is greater).

  11.1.5. For each Recovered fragment, using as many measurements as practical, calculate the mean concentration for each element.

  11.1.6. For each element, compare the mean concentration in the Recovered fragment to the match interval for the corresponding element from the Known fragments.

  11.1.7. If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not "match" and the glass samples are considered distinguishable.

- **This is a statistical inference procedure**

Statistical Inference - Background

- Definition - a **parameter** is a numerical characteristic of the population, e.g., a population mean μ, a population proportion π.
- Statistical methods are usually concerned with learning about population parameters from sample data.
- A key point - the mean of a sample and the mean of a population are different concepts.
- Idea: apply laws of probability to draw inferences from a sample:
  - Compute sample mean.
  - If we have a "good" sample, then this should be close to the population mean.
  - The laws of probability tells us how close we can expect them to be.

To re-iterate

- In statistics, the word *parameter* always refers to a population attribute.
- The value of a parameter is unknown, unless we measure the attribute in every population item and calculate it.
- We draw a sample from a population so that we can use *sample estimates* to infer the value of population parameters.

Statistical Inference - Background

- Goal: inference about a parameter.
- Possible parameters:
  - Mean concentration of aluminum in population of glass fragments from a given source.
  - Proportion of bags containing illicit substances in a large seizure.
  - Variability in the number of CMS in pairs of bullets fired from the same guns.

Some Common Notation

- Suppose that we draw a random sample of size n from a population.
- We measure attribute X on each sampled item and obtain $\{x_1, x_2, x_3, ...x_i, ..., x_{n-1}, x_n\}$ values.

Population parameter	Sample estimate	Calculation
Mean μ	$\bar{x}$	$(\sum_{i=1}^{n} x_i)/n$
Variance σ^2	S^2	$(\sum_{i=1}^{n}(x_i - \bar{x})^2)/(n-1)$
Standard deviation σ	S	$\sqrt{S^2}$
Proportion π	p	Proportion of "successes" in sample.

Sometimes we use "hats" to denote estimates, e.g., $\hat{\mu}, \hat{\pi}, \hat{\sigma}$ etc. instead of $\bar{X}, p, S$.

Different Types of Inference

- We can make different kinds of inferential statements about any population parameter:
  - Estimate of parameter (point estimate)
  - Range of plausible values for parameter (interval estimate)
  - Test a specific hypothesis about the value of a parameter

Point Estimation

- An *estimator* is a rule for estimating a population parameter from a sample.
- An *estimate* is the resulting value.
- For example, suppose that we have a sample of five signatures with complexity values equal to:
  $4, 3.2, 4.8, 5, 2.5$.
- The **estimator** of the population mean is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- The **estimate**, on the other hand, is 3.9.

- An estimator for the population standard deviation is:

$$S = \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \right]^{1/2}.$$

- The estimate we compute from the sample above is:

$$\begin{aligned} S^2 &= \frac{(4-3.9)^2 + \ldots + (2.5-3.9)^2}{4} \\ &= \frac{4.48}{4} = 1.12, \end{aligned}$$

so the standard deviation S is $\sqrt{1.12} = 1.06$.

Properties of Estimators

- There are many possible estimators for any population parameter.
- Example - suppose we are interested in estimating the population mean or average.
- Three possible estimators are:
  - The mean of a random sample from the population. (Spoiler alert: it is a very good estimator)
  - The median of a random sample is an alternative. (Less sensitive to wild measurements).
  - 47 is another possible estimator (A terrible estimator – unless we are very lucky!)
- How do we select an estimator from among all possible estimators?

Properties of Estimators

- Evaluate estimator by considering certain properties:
  - **Bias** - how close *on average* to population value.
  - **Variability** - how variable is the estimate from sample to sample.
- Before we talk about these properties, we need to discuss the concept of *sampling distributions*.

Sampling Distributions

- When we draw a random sample from a population, we could end with any one of a large number of possible samples.
- Recall that when sampling with replacement, the number of possible samples of size n from a population of size N is N^n.
- Suppose that the population consists of 100 balls labeled 1 through 100:
  - $\mu = 50.5$, $\sigma = 29.0$.
  - We draw $M = 2$ random samples of size $n = 5$ from the population.
- The table shows two potential samples we might obtain.

	Sample values	$\bar{x}$	S
Sample 1	47, 72, 43, 32, 85	55.8	21.9
Sample 2	1, 53, 63, 11, 65	38.6	30.3

- We notice four things (at least)
  - The two sample means and sample SDs are different, and different from the population parameters.
  - Sample estimates are *random* because like the sample items themselves, they are unknown until we actually draw the sample.
  - The mean of the two sample means is 47.2, which is closer to the true mean (50.5) than either of the two sample means.
  - The SD of the two sample means is 12.16, smaller than the true SD *of the population items* which was 29.0
- Now we draw $M = 10$ samples of size $n = 5$ and compute the 10 sample means. We get:
  55.8, 43.2, 56.8, 48.0, 43.0, 53.4, 55.8, 44.0, 55.0, 38.6.
- The mean of the means is now 49.44, quite close to the true mean $\mu = 50.5$

Sampling Distributions

- A **sampling distribution** is the distribution of a sample statistic (e.g., a mean or a standard deviation) over an infinite number of samples.
- If we have a sample of size n $\{x_1, x_2, ..., x_n\}$ from a population with mean μ and variance σ^2, then:
  - On the average, the mean of the n sample means is close to μ.
  - The variance of the sample means is σ^2/n.
- More formally, the sampling distribution of the sample means has expected value μ and variance σ^2/n.
- In the example above, the mean of the 10 means is 49.44, very close to the true mean 50.5.

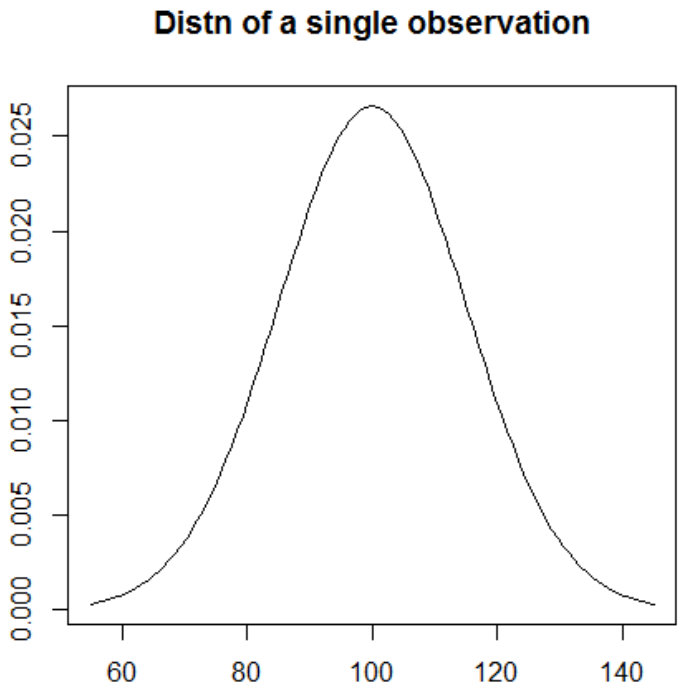
Statistical Inference - Standard Errors

- A limitation of just providing a point estimate is that it doesn't provide any indication of uncertainty.
- The *standard error* (SE) of an estimator measures the uncertainty in our estimate.
- The SE of the sample mean has already been introduced and is just the square root of the sampling variance, and can be estimated as:

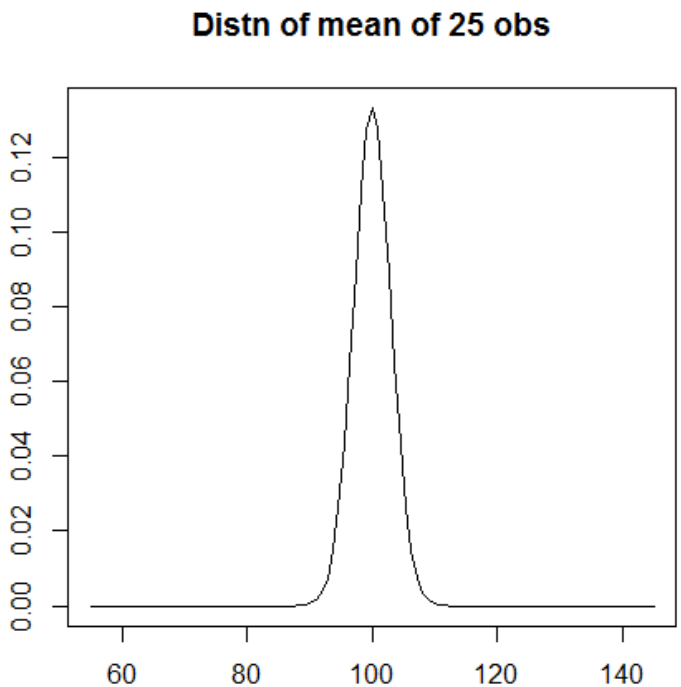
$$SE = \sqrt{\left(\frac{S^2}{n}\right)},$$

  with S^2 is the estimated population variance.
- All other sample estimates – SDs, percentiles, min, max, range – are also subject to uncertainty.
- SEs of statistics other than the mean are more difficult to calculate, but there are programs to do so.

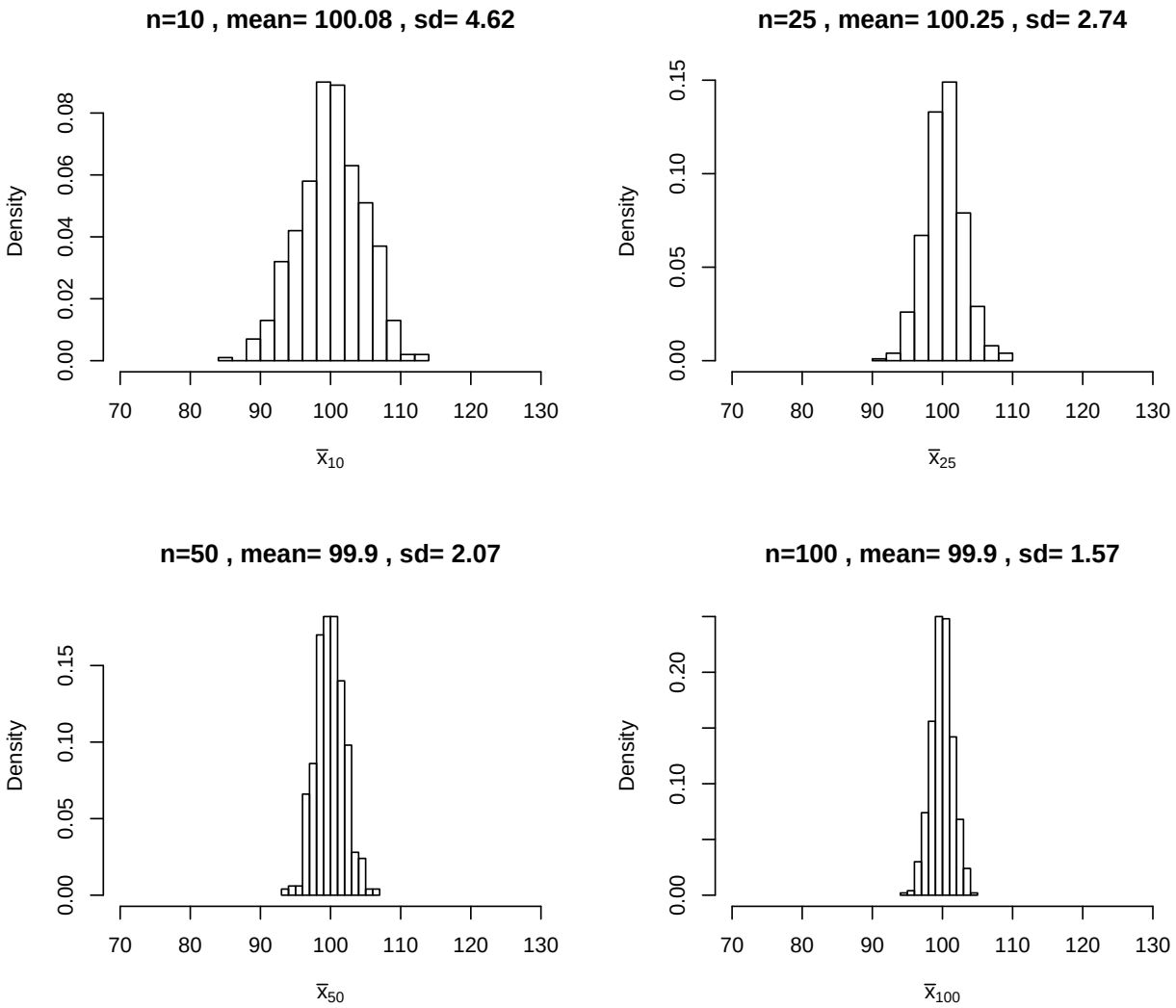
- Consider a normally distributed population with mean 100 and SD 15.
- We expect 95% of observations to be between 70 and 130: mean $\pm$ 2SD.



- Draw samples of size 25. The SE for the mean of 25 responses is 3 (SD divided by square root of sample size).
- Mean of this distribution is still 100.
- In 95 of 100 samples, the mean will be between 94 and 106.



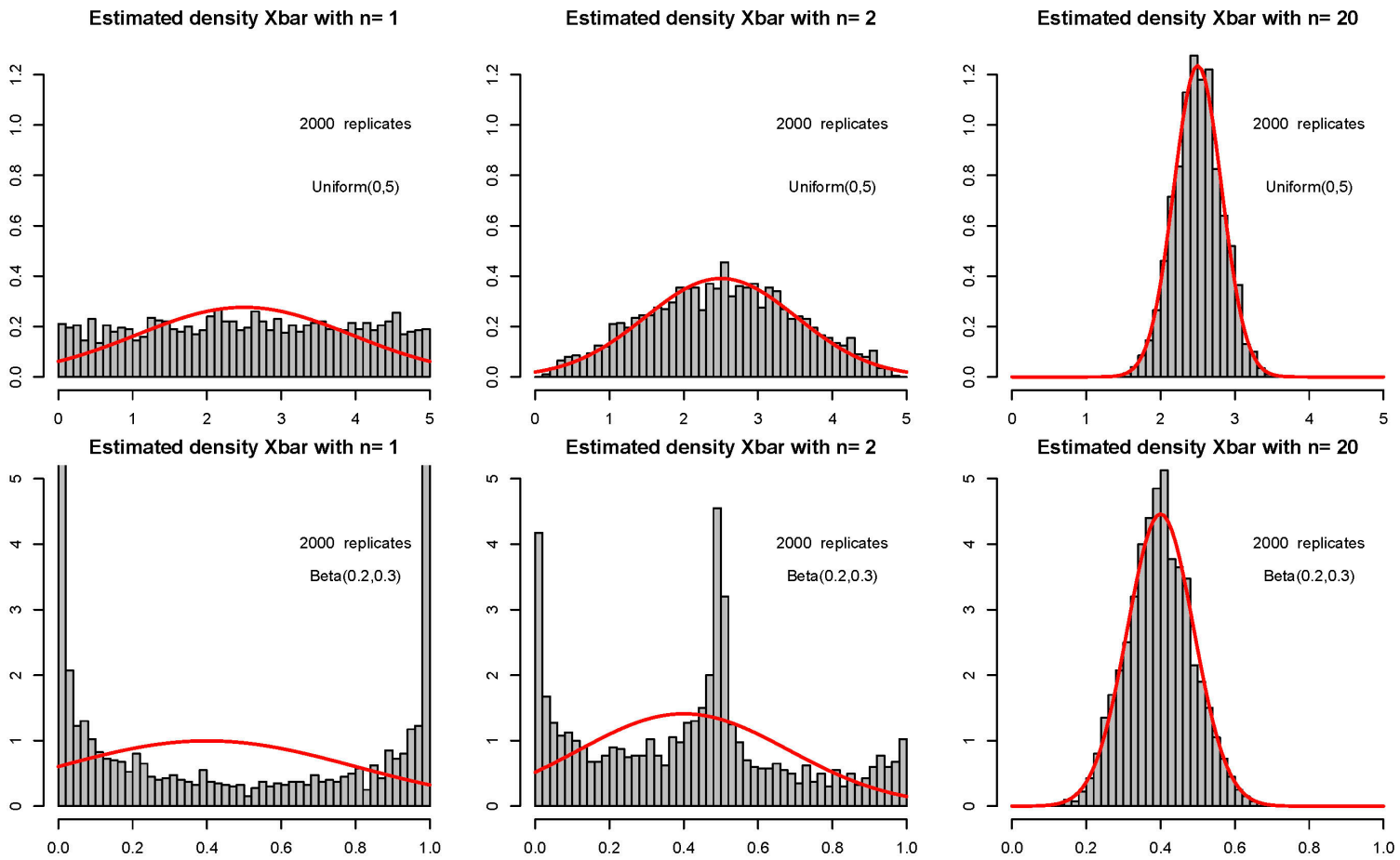
Effect of sample size.



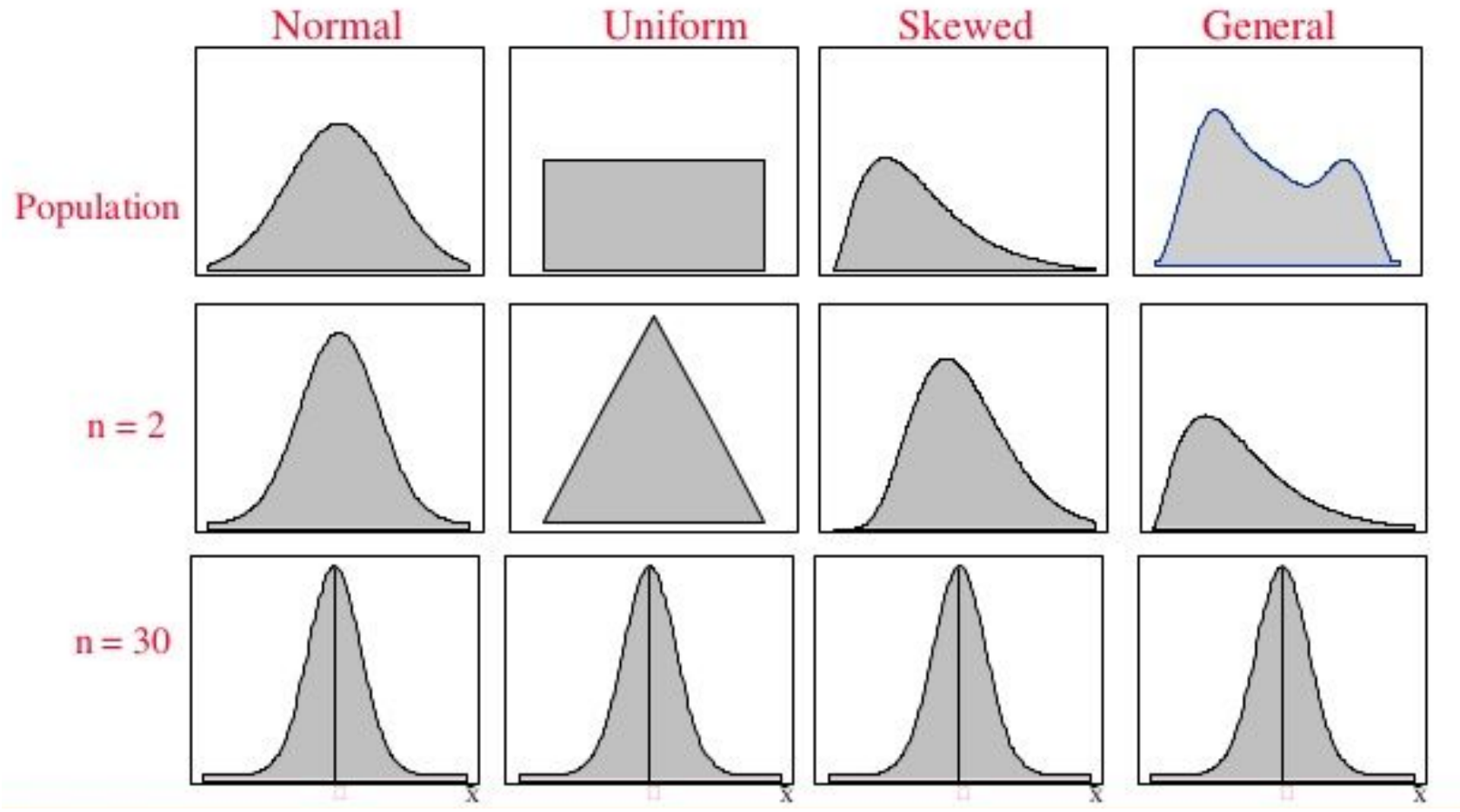
The Central Limit Theorem

- The CLT is a powerful theorem.
- It says that when we have large enough samples from a distribution with mean μ and variance σ^2, the distribution of $\bar{x}$ is normal with mean μ and variance σ^2/n *even when the distribution from which we sampled is not normal*.

CLT – Fig. 1


Estimated density Xbar with n= 1
2000 replicates
Uniform(0,5)
Estimated density Xbar with n= 2
2000 replicates
Uniform(0,5)
Estimated density Xbar with n= 20
2000 replicates
Uniform(0,5)
Estimated density Xbar with n= 1
2000 replicates
Beta(0.2,0.3)
Estimated density Xbar with n= 2
2000 replicates
Beta(0.2,0.3)
Estimated density Xbar with n= 20
2000 replicates
Beta(0.2,0.3)

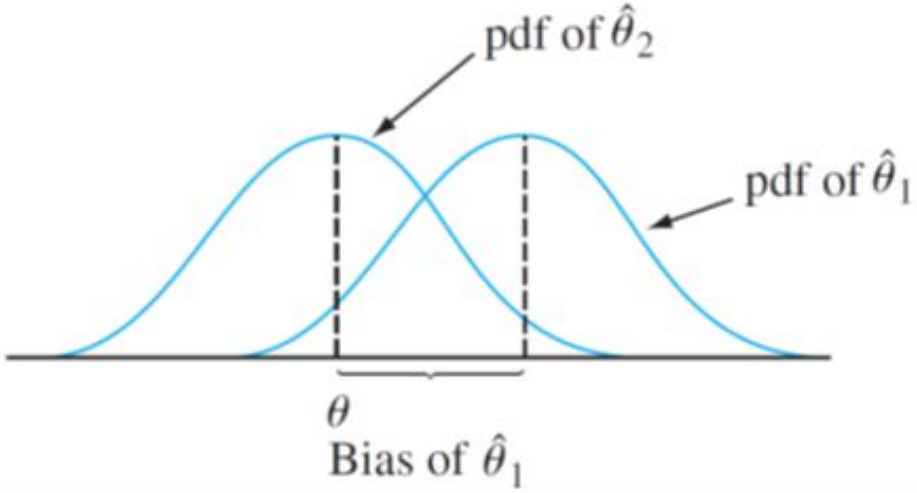
CLT – Fig. 2

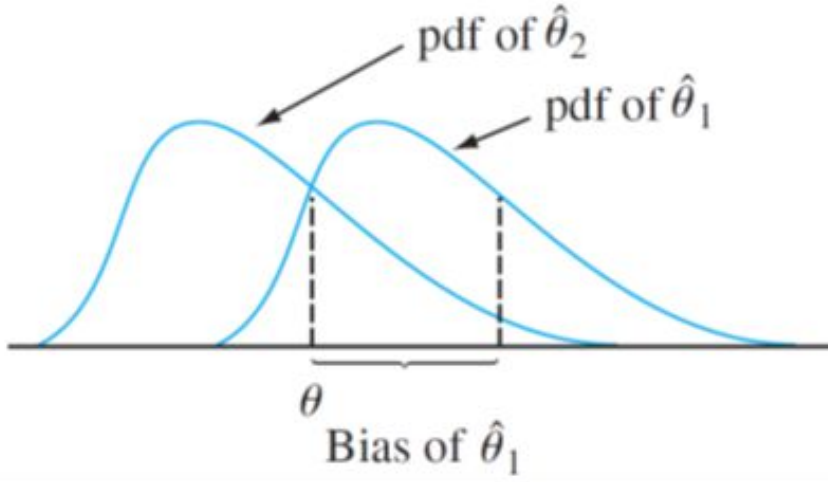


Back to Properties of Estimators

- Two results from the CLT is that the sample mean $\bar{x}$ has two good properties:
  - The sample mean is an **unbiased** estimator of the sample mean. Formally:
$$E(\bar{x}) = \mu.$$
  - The sample mean is a **consistent** estimator: as n increases, the sampling distribution of $\bar{x}$ gets more concentrated around μ.

Unbiasedness

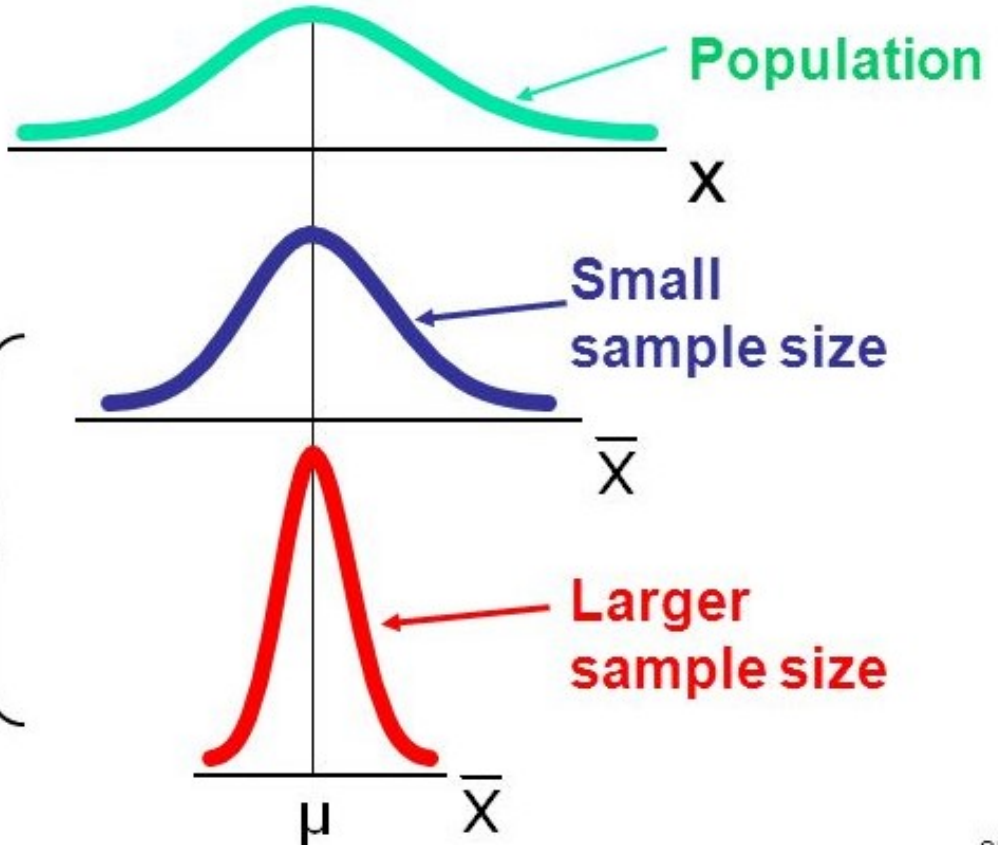




Consistency - Fig. 1

- The sample mean is a consistent estimator

(the value of $\bar{x}$ becomes closer to μ as n increases):



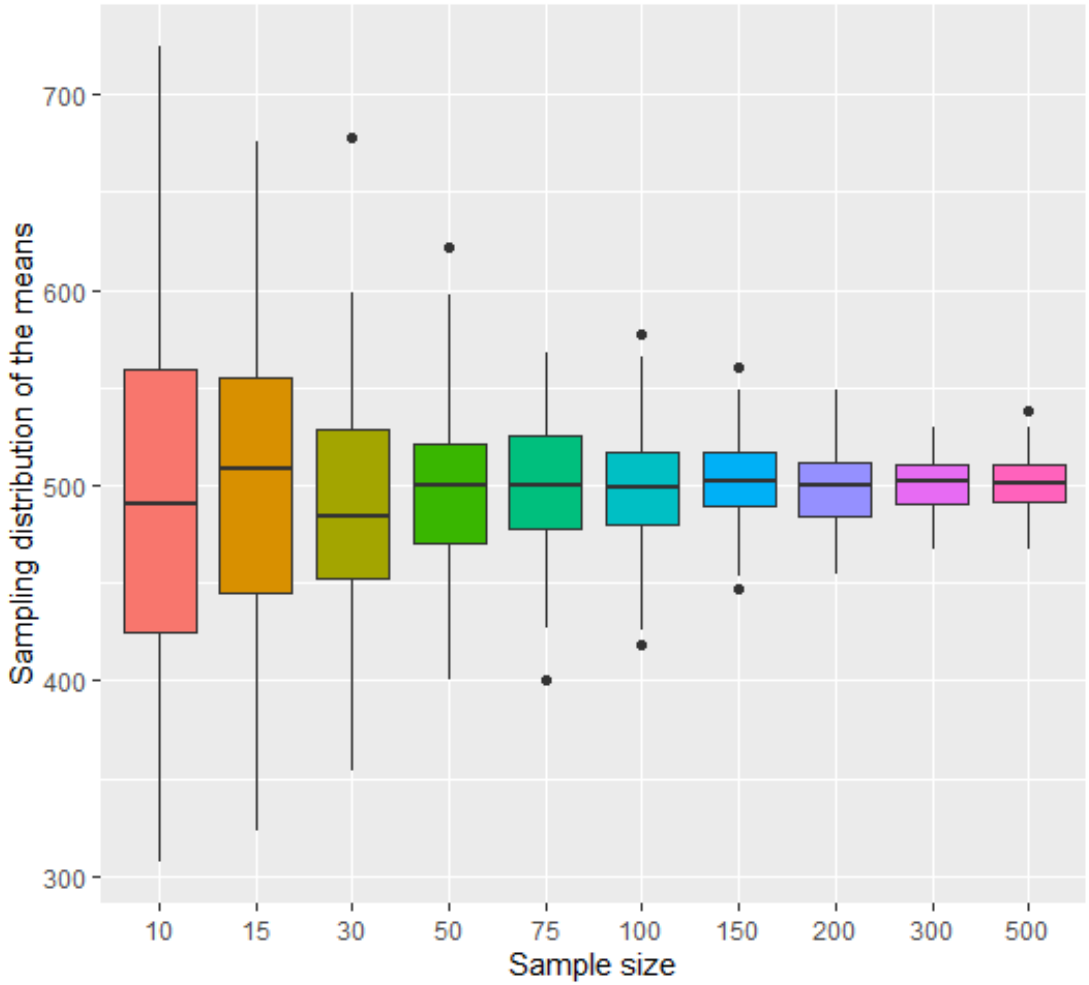
As n increases,

$\sigma_{\bar{x}} = \sigma/\sqrt{n}$ decreases

Bias/Consistency Simulation

- Population: 1000 items labeled 1 to 1000.
- Drew samples, with replacement, using values of n between 10 and 500.
- For each sample size, drew 100 independent samples.
- Box plots show the sampling distributions of $\bar{x}$ as n increases.

Consistency - Fig. 2



Statistical Thinking for Forensic Practitioners

Part 6: Inference (Video 2)

Dr. Alicia Carriquiry

Interval Estimation

- A confidence interval is an interval based on sample data that contains a population parameter with some specified confidence level.
- Essentially a confidence interval is a point estimate with additional information about its uncertainty.
- Typically we get an approximately 95% confidence interval for a quantity by taking point estimate $\pm$ 2 SEs (often use 1.96 which is correct for large samples from the normal distribution).
- In the case of estimation of the population mean μ:
  - Natural point estimate is the sample mean.
  - Approximate 95% confidence interval is

$$\text{sample mean} \pm 2 \text{ standard error}$$

.

Interval Estimation - Example

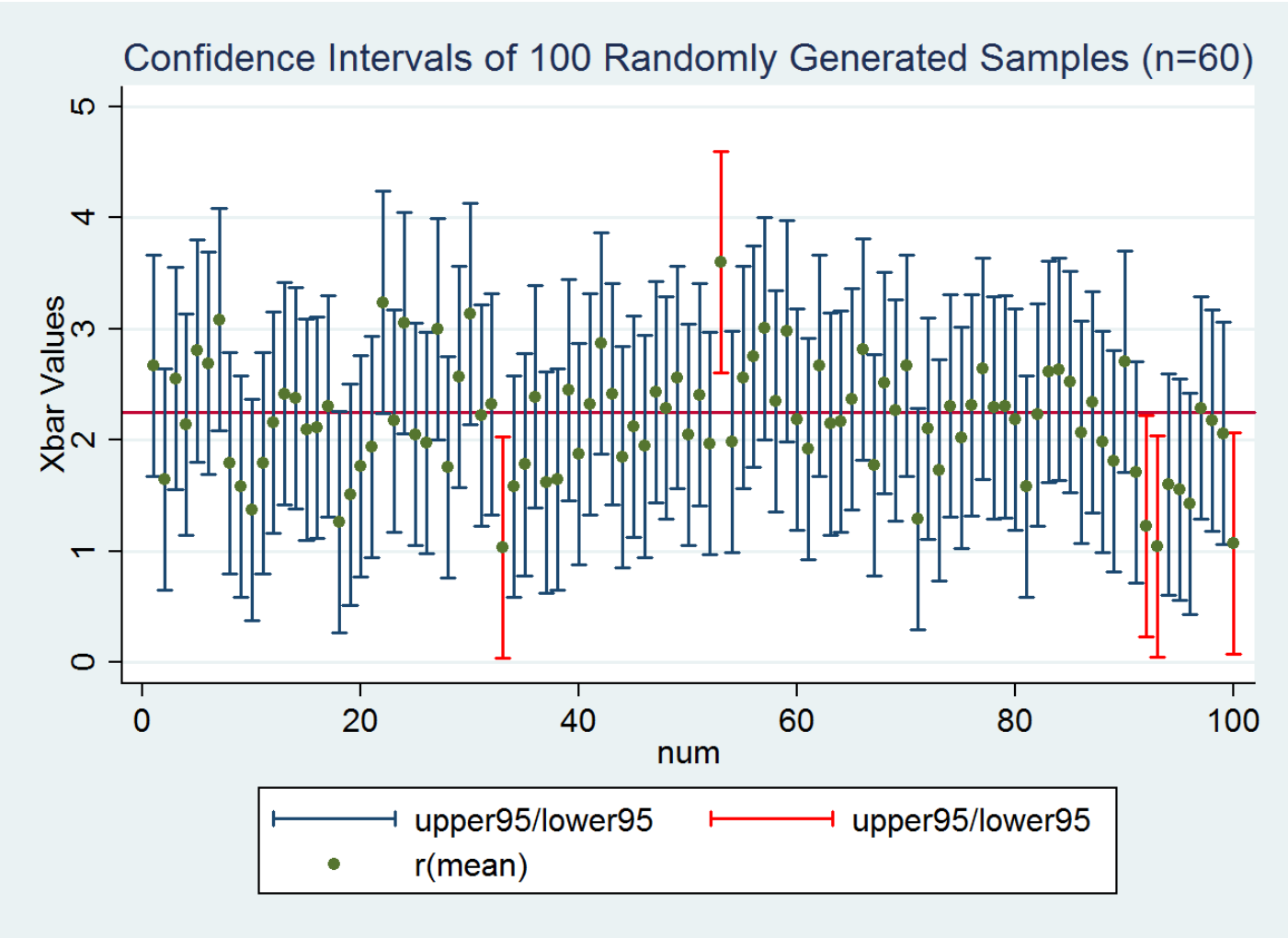
- Example: A window is broken in a house burglary case.
- CSIs gather 10 glass fragments from crime scene.
- Measure concentration of aluminum.
- Mean $= 0.730$, SD $= 0.04$.
- SE $= 0.040$ / sqrt(10) $= 0.013$.
- Approximate 95% confidence interval for the mean aluminum concentration in the crime scene window is

$$0.73 + / - 2 * 0.013 = (.704, .756).$$

Interpreting the Confidence Interval

- Like the sample itself, a CI is random.
- If we were to draw 100 samples from a population, each of the 100 intervals will differ from the others.
- Interpretation of confidence interval is important - 95% of intervals built in this way will contain the true population parameter.
- Note this type of interval (with higher confidence) is sometimes used in the analysis of glass evidence (ASTM 2926)

CI Simulation



A few observations

- The width of a confidence interval is:
  - Directly proportional to the confidence level we desire; everything else being equal, a 99% CI is wider than a 95% CI.
  - Also directly proportional to the variability of the measurements S^2. The more variable the observations, the wider the interval.
  - Inversely proportional to sample size; the larger the sample, the narrower the interval.
- A wide interval is less useful than a narrower one, because it suggests a higher level of uncertainty.
- For example, knowing that the victim died 24 hours $\pm$ 20 hours is not very useful, but knowing that the person died 24 hours $\pm$ 2 hours might be.

Estimating a Population Proportion

- Sometimes we wish to estimate a population proportion π (sometimes we use θ or just p).
- Examples might be:
  - Proportion of criminal cases in TX that involve a firearm.
  - Probability that a forensic professional in TX holds an MSc degree.
- As we did in the general case, we estimate a population proportion using a sample statistic, which we denote $\hat{\pi}$ or $\hat{p}$.
- If n is the sample size and Y is the "number of successes",:

$$\hat{\pi} = \hat{p} = \frac{Y}{n}.$$

Properties of $\hat{p}$

- Recall that the variable Y is distributed as a Binomial random variable with probability of success π.
- We have that:
  - $E(Y) = n\pi$, and $var(\pi) = n\pi(1-\pi)$.
  - Therefore:

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{n\pi}{n} = \pi,$$

$$Var(\hat{p}) = Var\left(\frac{Y}{n}\right) = \frac{n\pi(1-\pi)^2}{n} = \frac{\pi(1-\pi)}{n}.$$

- We note that the estimated proportion is both unbiased and consistent.

- An approximate 95% CI for π is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

  where:
  - We could have used 2 instead of 1.96.
  - We plug in $\hat{p}$ in the expression for the SE.
- The quantity to the right of the $\pm$ sign is called the *margin of error* or ME.
- Note that the ME depends on n.

Sample Size Calculation

- Suppose I wish to estimate a population proportion π.
- How big a sample do I need to achieve a $ME \leq 0.1$ with a confidence level of 95%?
- We work backwards, starting from the ME. If we fix the ME at 0.1, then:

$$\begin{aligned} 0.1 &= 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ 0.1^2 &= 1.96^2\frac{\hat{p}(1-\hat{p})}{n} \\ n &= \left(\frac{1.96}{0.1}\right)^2 \hat{p}(1-\hat{p}). \end{aligned}$$

- The sample size depends on what we think π might be.

Calculated Sample Sizes

p	1-p	ME	z	n
0.1	0.9	0.1	1.96	**35**
0.2	0.8	0.1	1.96	**61**
0.3	0.7	0.1	1.96	**81**
0.4	0.6	0.1	1.96	**92**
0.5	0.5	0.1	1.96	**96**
0.6	0.4	0.1	1.96	**92**
0.7	0.3	0.1	1.96	**81**
0.8	0.2	0.1	1.96	**61**
0.9	0.1	0.1	1.96	**35**

If π is 0.3 and the sample size is at least 81, my sample estimate $\hat{p}$ will be within 0.1 of the true π in 95 out of 100 samples.

Sample Size for Estimating μ

- The calculation is similar, but instead of the ME we often use a *relative* (RME) as the target.
- The RME is just the ME divided by μ, and can be set to 5% or 10% or any other percentage of the mean.
- Suppose you want to estimate the mean refractive index (RI) of float glass produced during a certain period of time in a certain plant.
- We wish to estimate the mean RI within 10% of its true value, so that $RME = 0.1 * \mu$.

Sample Size for Estimating μ

- So if we write the expression for the RME and solve for n:

$$\begin{aligned} \frac{ME}{\mu} &= \frac{1.96}{\mu}\sqrt{\frac{\sigma^2}{n}} \\ n &= \left(\frac{\sigma}{\mu}\right)^2 \left(\frac{1.96}{0.10}\right)^2. \end{aligned}$$

- Suppose that we think that the true RI is about 1.52 with variance 2.
- If we plug in the assumed values for μ, σ, the desired sample size is $n = 333$.
- If we wish to double the precision (i.e., be within 5% of the mean) then the required sample size is multiplied by 4.

Hypergeometric sample size calculation

- Sometimes the question is more complicated.
- Situation: We seize a shipment of teddy bears, which we believe may be concealing drugs.
- There are N teddy bears in the seizure. How many carry drugs?
- We want a sample of size n such that we can say that at least k% of the bears have drugs with 100(1-α)% confidence.
- E.g., if $k = 85, 0.05$ we want to open the smallest number of bears n so that we can be 95% confident that at least 85% of the bears carry drugs.
- n will have to be larger for larger k and smaller α.

Hypergeometric sample size calculation

- Recall the hypergeometric distribution from Part 4.
- If X has a hypergeometric distribution, then

$$\text{Prob}(X = k) = \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}}.$$

- We can choose k, α and we know N.
- By using expression above, we can solve for the n so that at least $K = kN$ are "successes" with confidence 100(1-α)%.
- There is software available to do these calculations.
- Two different situations:
  - When all n sample items are successes.
  - When one or more sample items are "failures" (e.g., teddy bear does not carry drugs).

All sample items positive

Population size N	95% confidence			99% confidence		
	k=0.5	k=0.7	k=0.9	k=0.5	k=0.7	k=0.9
10	3	5	8	4	6	9
20	4	6	12	5	9	15
30	4	7	15	6	10	20
40	4	7	18	6	10	23
50	4	8	19	6	11	26
60	4	8	20	6	11	28
70	5	8	21	7	12	30
80	5	8	22	7	12	31
90	5	8	23	7	12	32
100	5	8	23	7	12	33
200	5	9	26	7	13	38

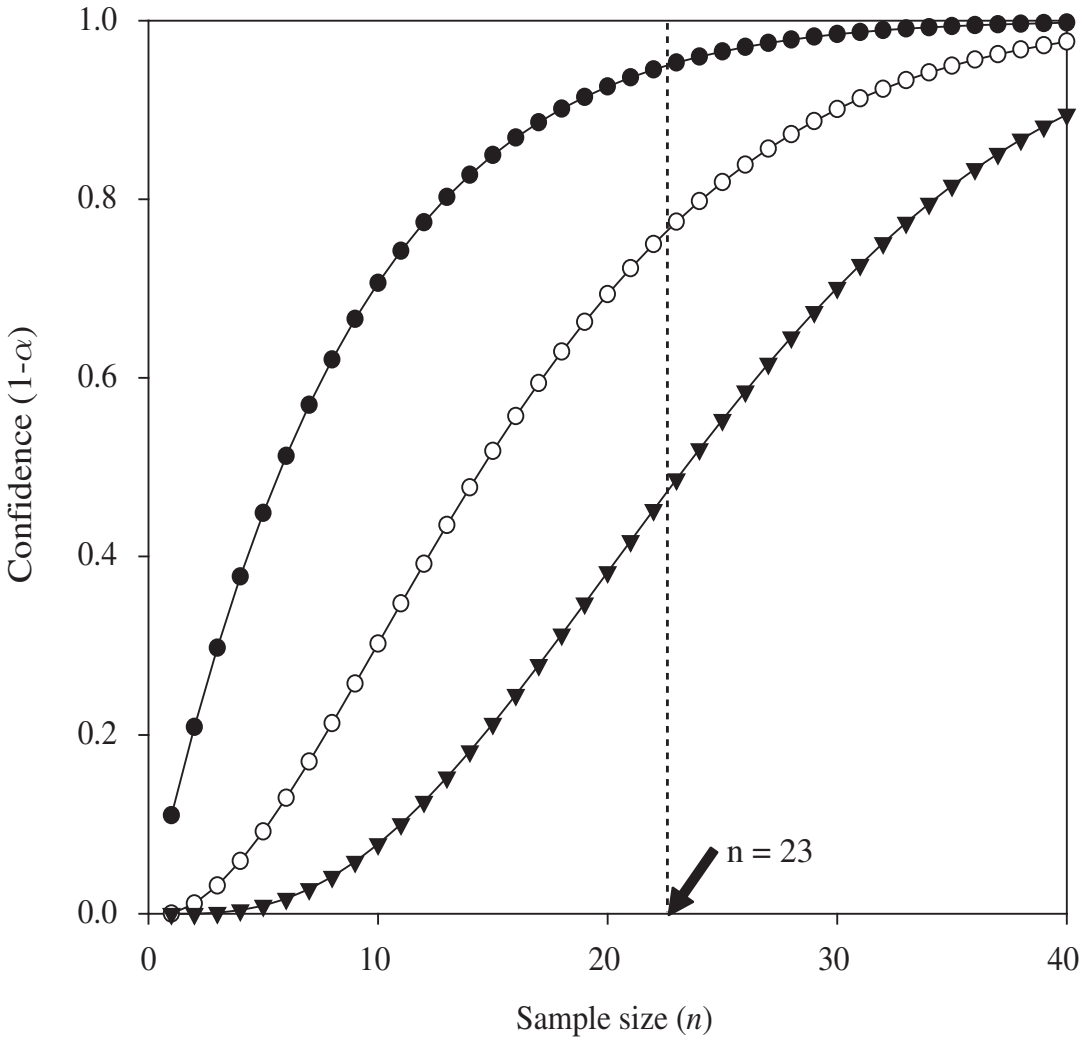
All sample items positive

Population size N	95% confidence			99% confidence		
	k=0.5	k=0.7	k=0.9	k=0.5	k=0.7	k=0.9
300	5	9	27	7	13	40
400	5	9	27	7	13	41
500	5	9	28	7	13	41
600	5	9	28	7	13	42
700	5	9	28	7	13	42
800	5	9	28	7	13	42
900	5	9	28	7	13	43
1 000	5	9	28	7	13	43
5 000	5	9	29	7	13	44
10 000	5	9	29	7	13	44

Some sample items negative

Population size N	95% confidence						99% confidence					
	k=0.5		k=0.7		k=0.9		k=0.5		k=0.7		k=0.9	
	1 neg	2 neg	1 neg	2 neg	1 neg	2 neg	1 neg	2 neg	1 neg	2 neg	1 neg	2 neg
10	5	7	7	9	10	–	6	7	8	9	10	–
20	6	8	10	13	17	20	8	10	12	14	19	20
30	7	9	11	14	22	27	8	11	14	17	25	29
40	7	9	12	15	26	32	9	11	15	18	30	35
50	7	10	12	16	29	36	9	12	16	20	34	41
60	7	10	12	16	31	39	9	12	16	20	38	45
70	7	10	13	17	32	41	10	12	17	21	40	48
80	7	10	13	17	34	43	10	12	17	21	42	51
90	7	10	13	17	35	45	10	13	17	21	44	54
100	7	10	13	17	36	46	10	13	17	22	46	56
200	8	10	14	18	40	53	10	13	18	24	54	67
300	8	10	14	19	42	55	10	13	19	24	57	71
400	8	11	14	19	43	57	10	13	19	24	58	74
500	8	11	14	19	44	58	10	14	19	24	59	75
600	8	11	14	19	44	58	10	14	19	25	60	76
700	8	11	14	19	44	59	11	14	19	25	61	77
800	8	11	14	19	44	59	11	14	19	25	61	77
900	8	11	14	19	45	59	11	14	19	25	61	78
1 000	8	11	14	19	45	59	11	14	19	25	62	78
5 000	8	11	14	19	46	61	11	14	20	25	64	81
10 000	8	11	14	19	46	61	11	14	20	25	64	81

Confidence vs sample size



Confidence against sample size ($N = 100$; $k = 0.9$) for 0, 1, and 2 negatives expected. Lines -●- for 0 negatives; -○- for 1 negative; -▼- for 2 negatives.

Statistical Thinking for Forensic Practitioners

Part 6: Inference (Video 3)

Dr. Alicia Carriquiry

Hypothesis Testing

- Sometimes we wish to formally test a hypothesis about a population parameter.
- The hypothesis to be evaluated is known as the null hypothesis and usually refers to an assumption of no difference or no change.
- The null hypothesis represents the proposition we put forth, and the idea is to challenge that proposition.
- To do that, we look for evidence against the null hypothesis.
- We formulate an alternative hypothesis that helps us to design the test.
- Notation:
  - H_0 is the null hypothesis.
  - H_a (or H_1) is the alternative.

Hypothesis Testing Steps

1. Formulate the two hypotheses.
2. Collect data, compute relevant sample statistics (e.g., mean and SE).
3. Calculate the "distance" between sample statistic and hypothesized parameter value.
4. Decide between the two hypotheses:
   - 4.1 Select a confidence level $(1 - \alpha)$.
   - 4.2 Determine *decision threshold* or *critical value* of the test.
   - 4.3 Compute a $p-$value.
   - 4.4 If the $p -$ value $\leq \alpha$ reject H_0 and conclude H_a.
5. Interpret results in the context of the original question.

Formulate the Hypotheses

- A simple test of hypothesis contrasts a point hypothesis against another point hypothesis, or more commonly, a point hypothesis versus the complement.
- For example, suppose that we wish to learn about the average width of duct tape made by a certain manufacturer.

  +6We postulate the the typical width of duct tape made by the company is 1.94 inches.
- The null here is: $H_0 : \mu = 1.94$.
- A potential alternative is: $H_a : \mu \neq 1.94$.
- This is a *two-tailed test*, because if we decide that the average width of tape is "significantly" below 1.94 or above 1.94, then we choose H_a over H_0.

One-sided Test

- Sometimes the alternative is one-sided.
- Example: bullets fired from different guns exhibit no more than 6 CMS. In this case, hypotheses are:

$$\begin{aligned} H_0: &\quad \text{CMS} < 6 \\ H_a: &\quad \text{CMS} \geq 6. \end{aligned}$$

- Intuitively, to decide between the two hypotheses small values of CMS will tend to confirm the null and large values will tend to confirm the alternative.

Let's Think About This Intuitively

- If we wish to test whether the population mean is equal to some value m or not, then intuition says that we should:
  - Obtain a sample from the population and compute the sample mean.
  - Compare the sample mean to m.
  - Decide in favor of H_0 if sample mean is "close enough to" m.

  Several questions arise:
  - How do I measure the distance between the sample mean and m?
  - What do I mean by "close enough to"?

Observation

- Remember that H_0 is formulated after you have done research on the subject, so it is NOT some wild guess.
- Thus, to reject H_0 we seek overwhelming evidence against it.

Collect Data

- The second step in a test of hypothesis is to draw a sample of size n from the relevant population, and:
  - Compute sample statistics such as $\bar{x}, S, SE$.
- Determine n by thinking about the desired accuracy:
  - If m is the hypothesized value of μ, how small a difference between $\bar{x}$ do I want to be able to detect?
  - Or, what is the smallest distance between $\bar{x}$ and m that would make me reject H_0?
  - This is called an *effect size*.
- Larger samples allow us to detect smaller effect sizes.

Distance between $\bar{x}$ and hypothesized value

- Basic idea of hypothesis testing is to compute a test statistic that measures 'distance' between the data we have collected and what we would expect under the null hypothesis.
- Intuitively, if $\bar{x}$ is the sample mean and m is the hypothesized value, then something along the lines of

$$\bar{x} - m$$

would seem to make sense.

- One problem is that a large difference in some applications would be a tiny difference in others, so it would be good to avoid this ambiguity by standardizing the difference in some way.
- We want a standardized difference that is independent of the scale of the measurements.

Distance between $\bar{x}$ and hypothesized value

- Typically we compute a statistic of the form

$$\frac{\text{point estimate - null hypothesis value}}{\text{SE of estimate}} = \frac{\bar{x} - m}{SE_{\bar{x}}},$$

  where $SE_{\bar{x}}$ is a standard error of $\bar{x}$.
- Can be interpreted as the number of standard errors the sample estimate is from the hypothesized value if the null hypothesis is true.
- Since the mean and the SE are in the same units, the test statistic is unit-less.
- When the test statistic is large (either in the negative or the positive directions), then we reject the H_0.
- *What do we mean by "large"?*

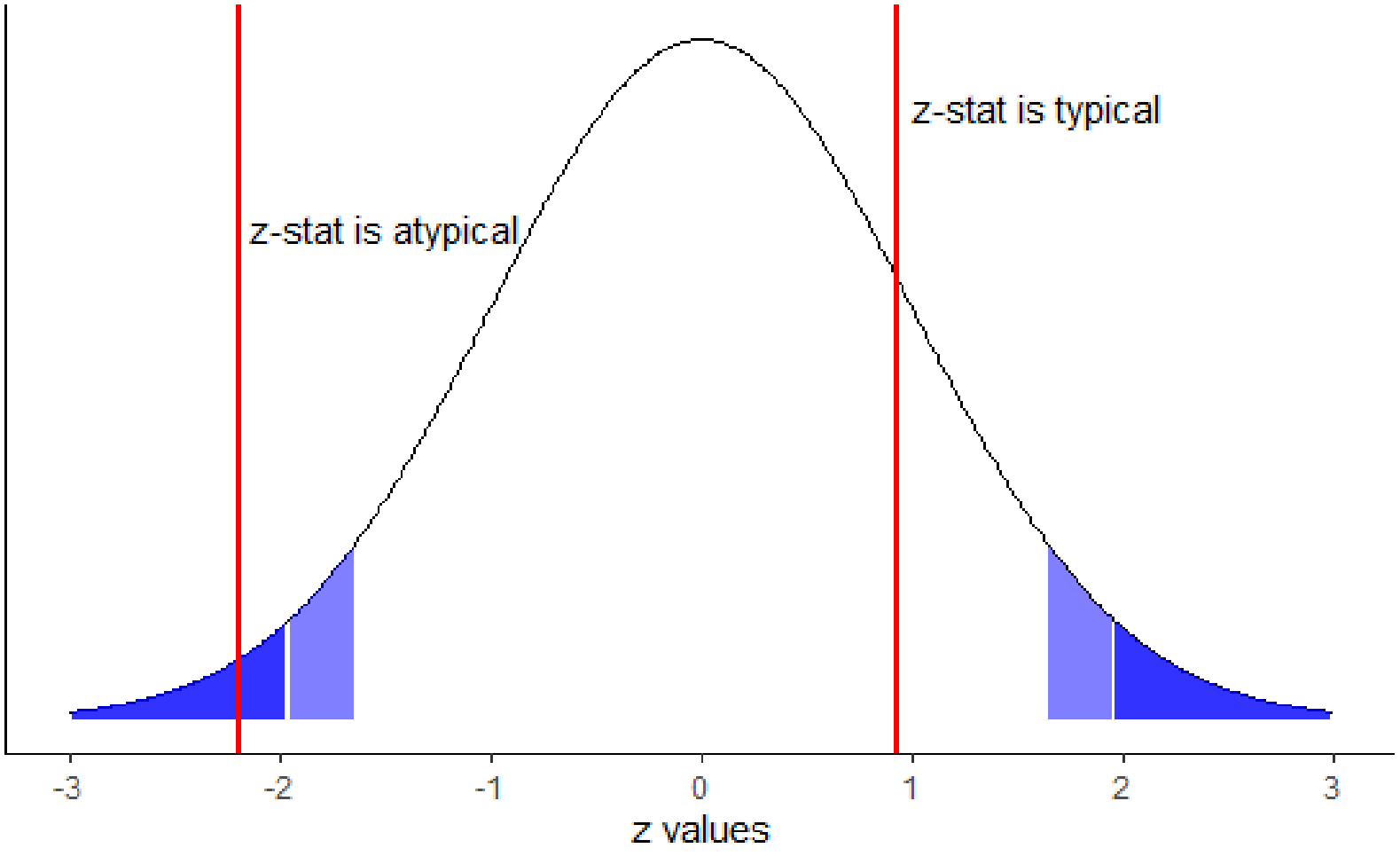
Interpreting the test statistic

- For a minute, assume that we know the population variance, so instead of the SE we can use $\sigma/\sqrt{n}$ in the denominator of the test statistic.
- Further, assume that H_0 is true.
- Under those assumptions:

$$z = \frac{\text{mean} - \text{hypothesized value}}{\sigma/\sqrt{n}} = \frac{0}{\sigma/\sqrt{n}},$$

  which is distributed as a normal random variable $N(0, 1)$.
- When H_0 is true, we expect to observe values of z between -2 and 2 with about 95% chance.
- We also expect to see extreme values of z below -2 or above 2 with about 5% chance.

"Typical" values of z



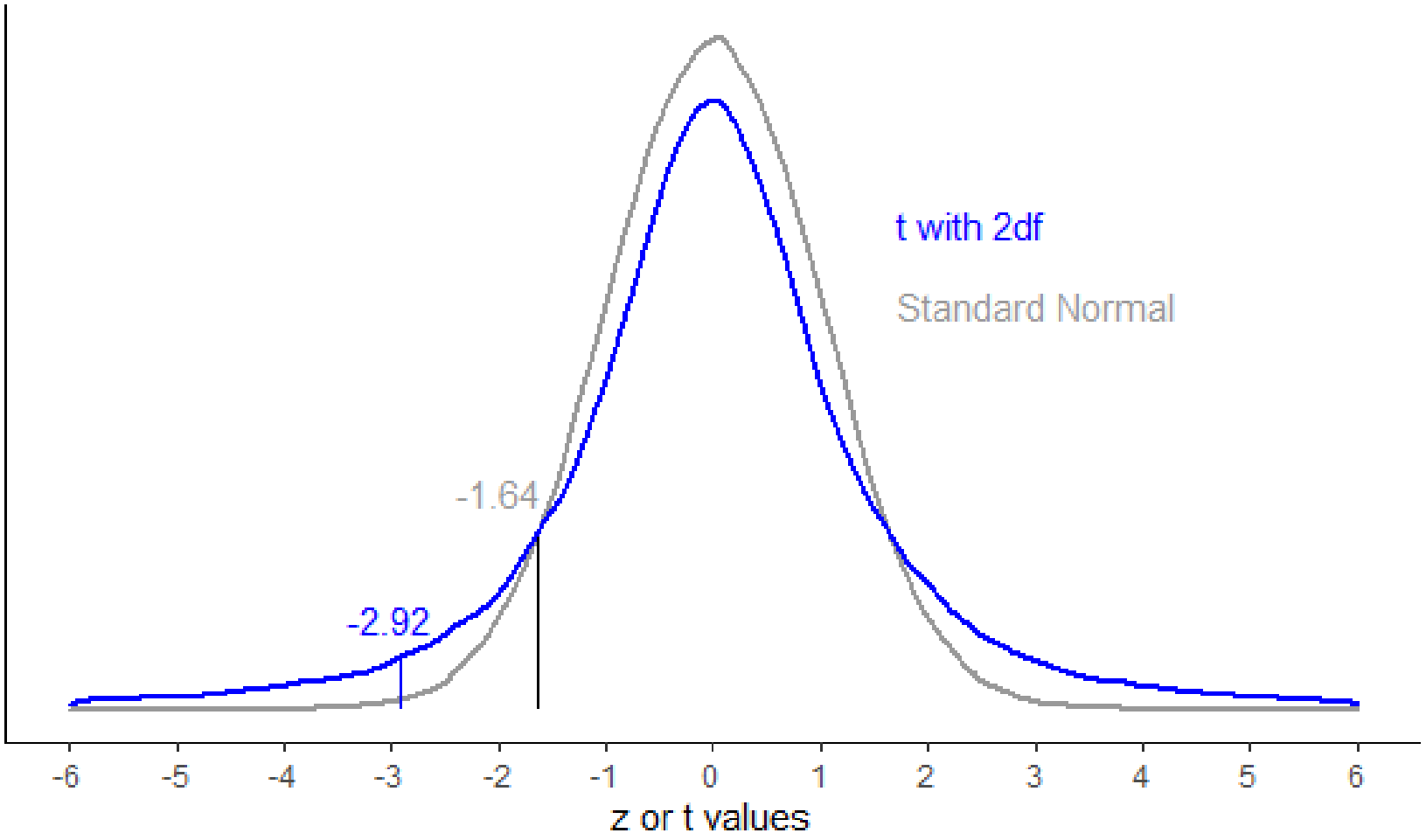
The t-statistic

- In realistic scenarios, we do not know σ^2 so we estimate it from the sample as S^2.
- The test statistic we compute is now:

$$t = \frac{\text{point estimate - null hypothesis value}}{\text{SE of estimate}},$$

  mentioned earlier, which is a random variable with a student-t distribution with $n - 1$ degrees of freedom.
- If H_0 is true, the mean of the t-statistic is also 0.
- For small n (below about 30), the t-distribution has "fatter tails" than the normal distribution:
  - To find 95% of the distribution I need to go a bit further outside -2 and 2.
  - The smaller the n, the larger the t test statistic needs to be before I conclude that the population mean is different from the hypothesized value.

t versus z Distributions

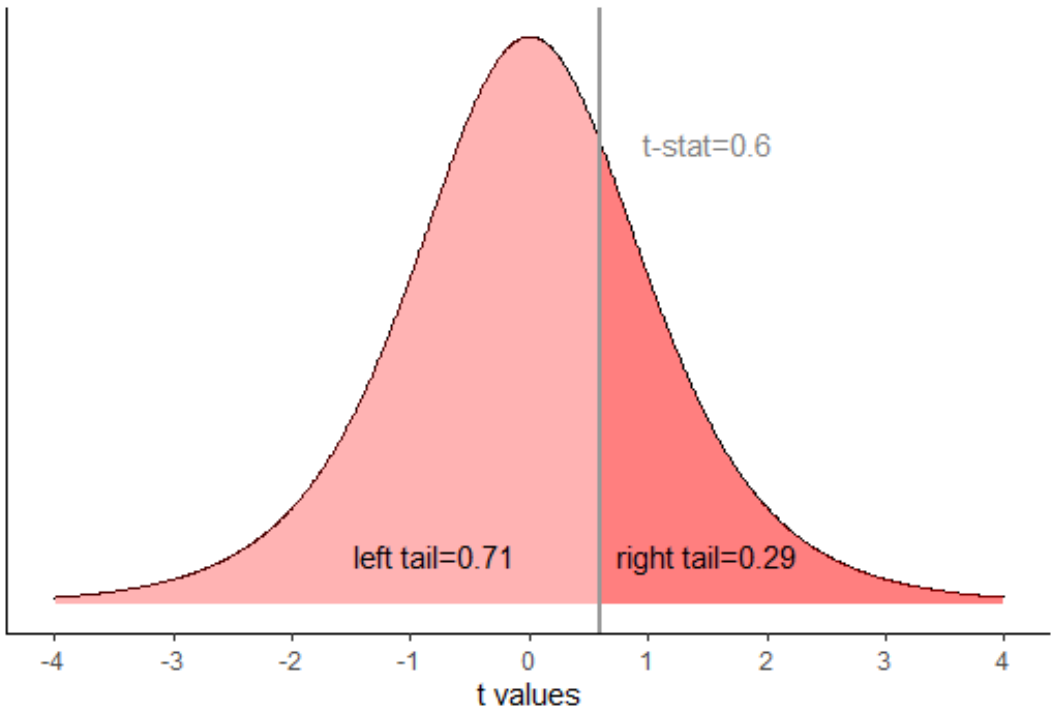


Mini-Example

- You have almost certainly seen t-tests in college or maybe in high school.
- In the old days, we computed the t-statistic and then looked in a t-table to see whether our statistic was *significant*.
- We will talk about this more in the next few slides, but just as a preview, looking up values in a table is no longer necessary.
- There are many programs that will do the "looking" for us, including Excel.

Mini-Example

- Example, n =6 (so df = 5) and t-statistic=0.6:
  - `=T.DIST(0.6, 5, TRUE)` results in 0.71, which is the probability of observing a $t-$value below 0.6.
  - The function `=T.DIST.RT(0.6, 5)` returns the right tail, or $1 - 0.71 = 0.29$.



Decision Time: Type I and Type II Errors

- At this point, we need to decide between the two hypothesis.
- When we choose between the two hypotheses, we can err in two ways:
  - Type I: reject the null hypothesis when it is true (false positive)
  - Type II: fail to reject the null when it is false (false negative)
- Type I error often considered more serious: we only want to reject the null if strong evidence against it.
- When we carry out a test of hypotheses, we select the type I error:
  - The lower the probability of a type I error, the harder we make it to reject H_0.
  - The type I error probability is the threshold against which we compare the p-value (see later).

Observations

- Note that these statistical testing ideas are closely related to concepts in the justice system.
  - Null hypothesis = innocent, alternative = guilty.
  - Type I error is to decide guilty when person is innocent.
  - Type II error is to decide innocent when person is guilty.
- The two types of errors are relevant to forensics as well, but we often use the false positive/negative terminology (or sensitivity/specificity).

Possible Decisions

		Reality	
		Positive	Negative
Study Finding	Positive	**True Positive** (Power) (1-β)	False Positive **Type I Error** (α)
	Negative	False Negative **Type II Error** (β)	**True Negative**

Confidence Level and Critical Value

- Suppose that my t-statistic is 2.1. Do I conclude H_0 or H_a?
- I need a confidence level and a critical value to decide.
- To choose a *confidence level* $1 - \alpha$, remember that:
  - α is the probability of a type I error.
  - If $\alpha = 0.05$ then I am willing to incorrectly reject H_0 with probability 0.05.
- What about β, the probability of a type II error? Requires that we do some sample size calculations (see later).
  - β depends on sample size, effect size, α.
- Often, we worry only about α and just strive to get a large enough sample to avoid a high β.
- The *power of the test* is defined as the probability of correctly rejecting H_0, or finding an "effect" or a "signal" even if it is small. Power is computed as $1 - \beta$.

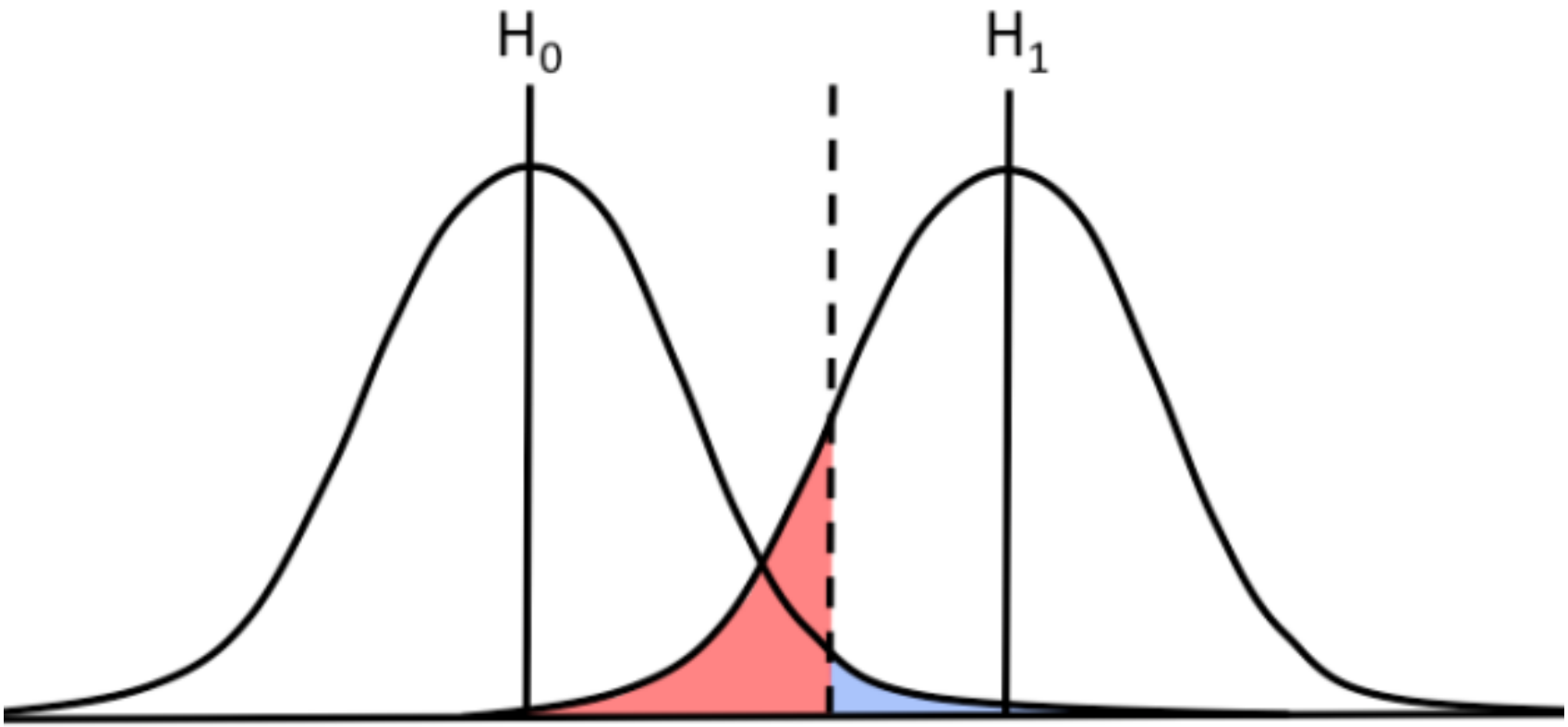
Critical Value

- Once I have decided on the value of α, I can determine the *critical value* for the test.
- The critical value is the decision threshold. If t^* is the critical value, then:
  - Reject H_0 if $t \leq -t^*$ or if $t \geq t^*$.
  - Smaller values of α correspond to larger values (in absolute value) of t^*.
  - Larger t^* means that it is more difficult to reject H_0: the difference between the sample mean and the hypothesized value has to be pretty large before we conclude that they are different.

Critical Values for the t-distribution

- For the t-distribution, the critical value depends on the df on α, and on whether the alternative is one or two-sided:
  - $\alpha = 0.05$, $df = 15$, two-sided: $t^* = |2.13|$ 0.025 probability in each tail.
  - $\alpha = 0.10$, $df = 15$, two-sided: $t^* = |1.75|$ 0.05 probability in each tail.
  - $\alpha = 0.05$, $df = 15$, one-sided: $t^* = -1.75$ or1.75 depending on whether H_a is $\mu \leq m$ or $\mu \geq m$.
  - $\alpha = 0.10$, $df = 15$, one-sided: $t^* = -1.34$ or 1.34.
- We can get the critical value using Excel:
  - `T.INV(0.025, 15)` gives -2.13.
  - `T.INV(0.975, 15)` gives 2.13.
  - `T.INV.2T(0.05, 15)` also gives 2.13. (This is the two-tailed version of `T.INV`).

Threshold, Type I, Type II Errors

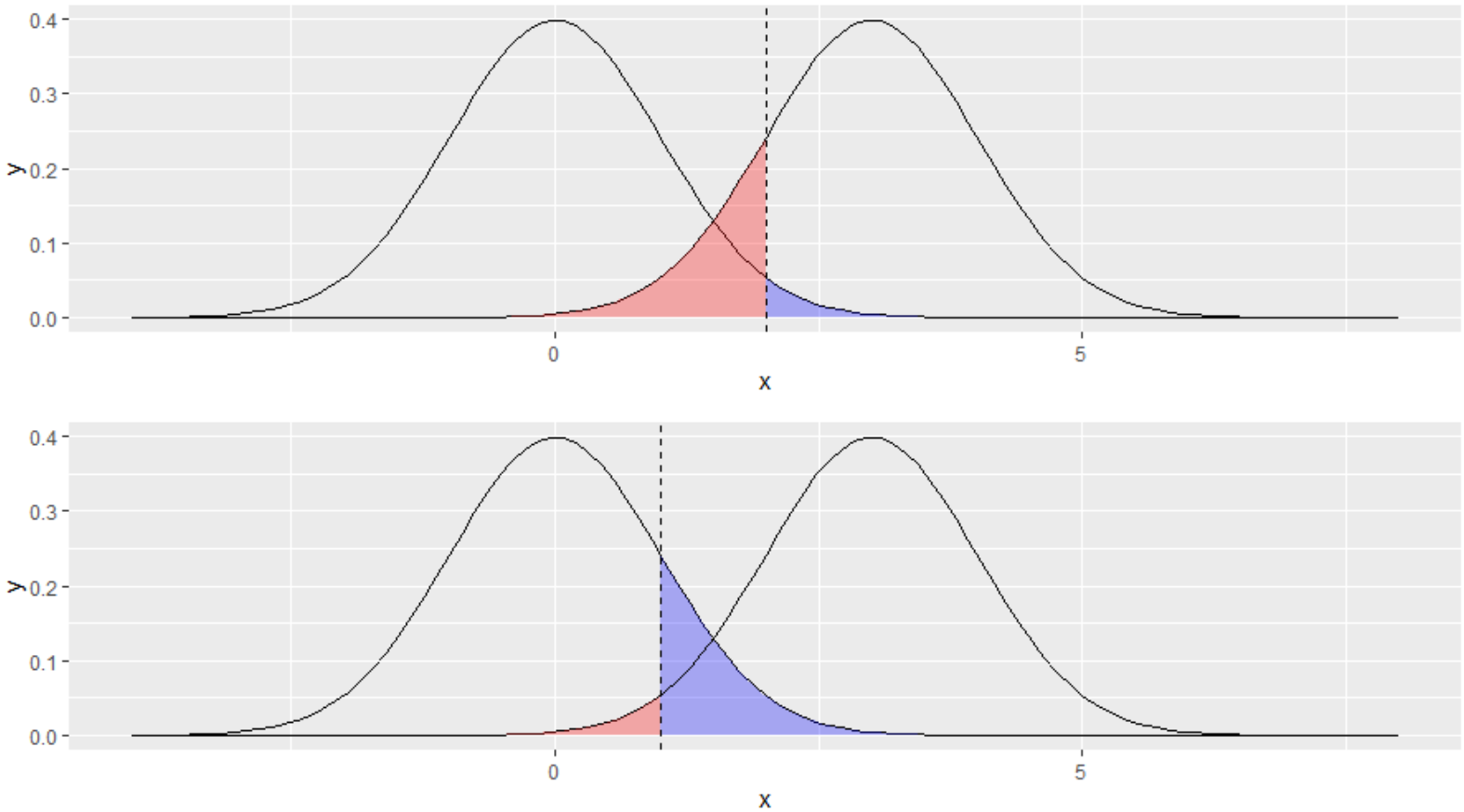


Strike a balance

Cannot minimize both errors at the same time.

Top panel: low type I, high type II.

Bottom panel: low type II, high type I.



Decision Time: $p-$values

- We can make a decision between the two hypotheses by comparing the $t-$statistic to the critical value:
  - If $t-$stat is more extreme than t^*, reject H_0.
- One drawback with that approach is that we do not know whether we rejected the null by a mile or by a hair.
- A common alternative is to attach a probability to the test statistic to quantify the "agreement" between the sample and H_0.
- Definition: a p-**value** gives the probability of getting data like the data we have observed (or something even more extreme) if the null hypothesis is true.
- Small p-values mean unusual data that lead us to question the null hypothesis (since sample data are unlikely to happen by chance).

Back to the Duct Tape Example

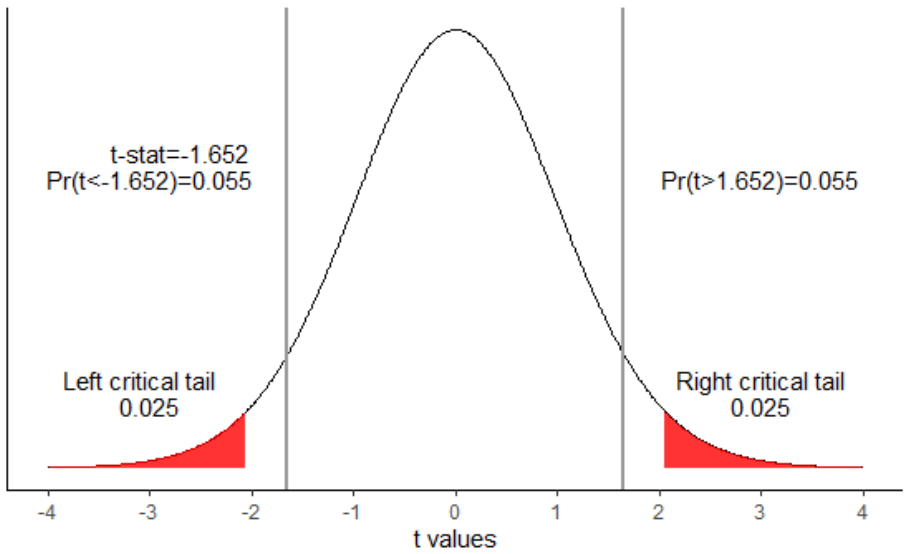
- We postulate that the mean width of tape made by company X is 1.94", so that $H_0 : \mu = 1.94$.
- $H_a : \mu \neq 1.94$, so test is two-tailed.
- We buy several rolls of tape and cut $n = 30$ random sections. Measure their width.
- Sample mean: $\bar{y} = 1.862$, sample SD: $S = 0.257$, standard error of the mean SE: $SE = 0.047$.
- The t-statistic is: $(1.862 - 1.94)/0.047 = -1.652$.

Continue Duct Tape Example

- The $p-$value is the probability of observing a t-statistic that is smaller than -1.652 or larger than 1.652 if H_0 is true.
- Since the t-statistic has a t-distribution with 29 df, we find that the $p-$value is:

$$\text{Prob}_{t_{29}}(t \leq -1.652) + \text{Prob}_{t_{29}}(t \geq 1.652) = 2 \times 0.055 = 0.109.$$

- Fail to reject H_0 at 95% confidence because $0.109 > 0.05$.



With a Bigger Sample...

- Suppose instead that we cut $n = 60$ sections of tape.
- Sample mean: $\bar{y} = 1.860$, sample SD: $S = 0.246$, standard error of the mean SE: $SE = 0.032$.
- The t-statistic is: $(1.860 - 1.94)/0.032 = -2.644$.
- Now the t-statistic has a t-distribution with 59 df, so we find that:

$$\text{Prob}_{t_{59}}(t \leq -2.644) = 0.0052,$$

  so the $p-$value is $2 \times 0.0052 = 0.0104$.
- Select a 95% confidence level, so that $\alpha = 0.05$.
- Since now $p-$value is less than 0.05, we reject H_0 and conclude H_a.

One-sided Example

- If instead we test $H_0 : \mu \leq 1.94$ versus $H_a : \mu > 1.94$, nothing changes except the critical region and the $p-$value.
- For $n = 30$, the t-statistic was -1.652.
- For $\alpha = 0.05$ the critical region is $t \leq -1.699$.
- Since the t-statistic is not in the critical region, we fail to reject H_0 by a hair!
- The $p-$value in this case is:

$$\text{Prob}_{t_{29}}(t \leq -1.652) = 0.055,$$

barely above α.

Statistical Thinking for Forensic Practitioners

Part 6: Inference (Video 4)

Dr. Alicia Carriquiry

Comparing Two Means

- In practice, we are often interested in comparing two samples (or more precisely, two populations).
  - Example 1: Do the glass fragments on the suspect originate from the broken window at the crime scene?
  - Example 2: were the bullets in the cadaver fired from the suspect's gun?
- In both cases, we have two populations, one from which crime scene items originate and the other from which suspect's items originate.
- The question of interest is whether those two populations are one and the same.
- This is the *same source* question.

Glass Example

- In the glass example, suppose that we have broken glass at a crime scene and glass fragments on the suspect.
  - Define μ_{scene} to be mean trace element level for the "population" of glass at the scene.
  - Define $\mu_{suspect}$ to be the mean trace element level for "population" of glass on the suspect.
- Is it plausible that glass fragments on suspect came from the crime scene (i.e., $\mu_{suspect} = \mu_{scene}$)?

Comparing Two Means - Steps

- Formulate the two hypotheses:
  $H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$.
  $H_a : \mu_1 \neq \mu_2$ or $H_0 : \mu_1 - \mu_2 \neq 0$,
  where population 1 is the crime scene population and population 2 is the suspect population.
- Obtain independent random samples from each of the two populations:
  $x_1, x_2, ..., x_{n1}$ from population 1.
  $y_1, y_2, ..., y_{n2}$ from population 2.
- Compute sample statistics:
  - $\bar{x}, S_x^2$ from sample 1.
  - $\bar{y}, S_y^2$ from sample 2.
- Intuition: If $\bar{x}$ and $\bar{y}$ are "close", we tend to believe H_0.

- As before, we calculate the t-statistic, which quantifies the difference between $\bar{x}$ and $\bar{y}$ in SE units:

$$t = \frac{(\bar{x} - \bar{y}) - 0}{SE_{diff}},$$

where 0 is the hypothesized value for the difference in population means and SE_{diff} is:

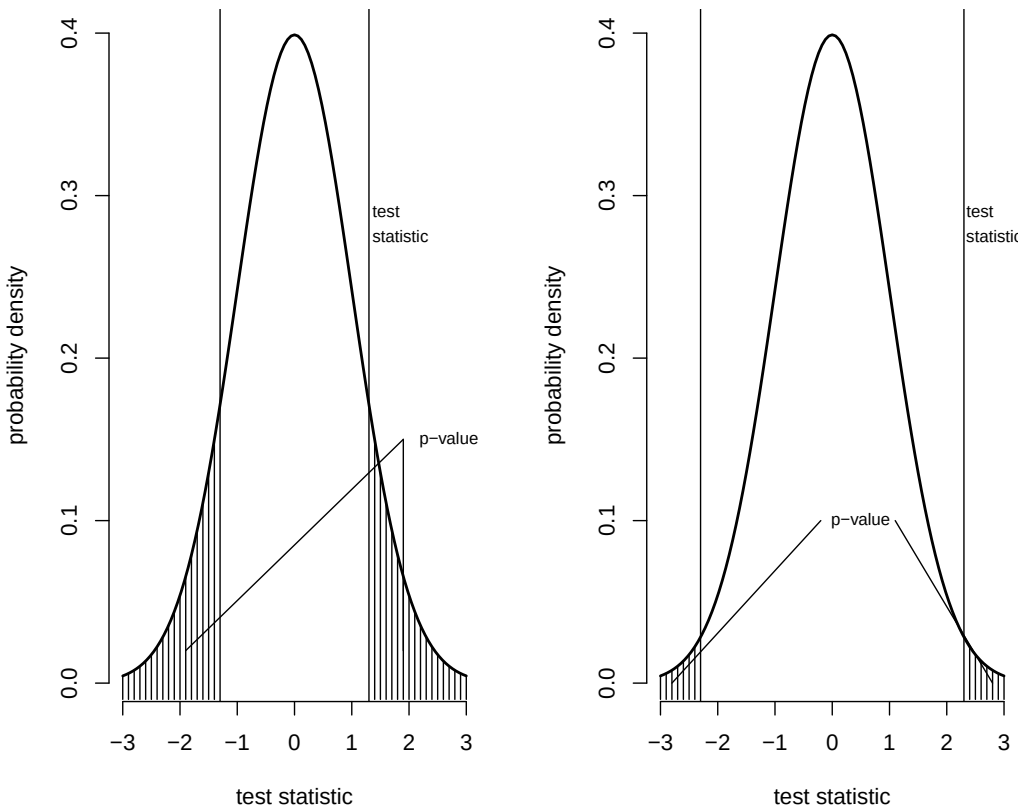
$$\begin{aligned} SE_{diff} &= [\text{Var}(\bar{x}) + \text{Var}(\bar{y}) - 2\text{Cov}(\bar{x}, \bar{y})]^{1/2} \\ &= [\text{Var}(\bar{x}) + \text{Var}(\bar{y})]^{1/2} \text{ (by independence)} \\ &= \left[\frac{S_x^2}{n1} + \frac{S_y^2}{n2}\right]^{1/2}. \end{aligned}$$

Hypothesis Testing - Compare Two Means

- We reject the null hypothesis if the t-statistic is large in absolute value.
- To decide what we mean by "large", we select a confidence level, as before, and compute a $p-$value for the test.
- The t-statistic for comparing two means has a t distribution, with $n1 + n2 - 2$ degrees of freedom.
- As before, the $p-$value is the probability of observing a more extreme t-statistic when the H_0 is true.
- Key statistical result is that these procedures work well even if populations are not normally distributed as long as the sample size is large. (The magic of the CLT!)

Hypothesis Testing - Normal Data

- Left figure: observed test statistic = 1.3, p-value = 0.19
- Right figure: observed test statistic = 2.3, p-value = 0.02



Comparing Two Means - Example

- Suppose $n1 = 10$ glass fragments are taken from glass at the scene (Y) and $n2 = 9$ fragments are found on the suspect (X).
  - $\bar{x} = 5.3, S = 0.9, SE(\bar{x}) = 0.9/\sqrt{10} = .285.$
  - $\bar{y} = 5.9, S = 0.85, SE(\bar{y}) = 0.85/\sqrt{9} = .283.$
  - observed difference is $\bar{y} - \bar{x} = 0.6.$
  - Standard error for this difference is $SE_{diff} = \sqrt{.285^2 + .283^2} = 0.402.$
- The test statistic is $0.6/0.402 = 1.5$ which yields a $p-$value of 0.15.
- If we had selected an $\alpha = 0.05$, or even 0.1, we fail to reject the hypothesis that the two glass populations are indistinguishable.

Interpretation of Results

- If the test fails to reject the null hypothesis, then we would say the two populations are indistinguishable.
- Interpretation is a key issue:
  - If we cannot reject the hypothesis of equal means, then we cannot *exclude* the possibility of a common source.
  - But we cannot conclude common source either!
- Much more about this in Part 8.

CI for Difference of Means

- As an aside, we have all the ingredients we need to compute a $100(1-\alpha)\%$ CI for the difference between two population means $\mu_1 - \mu_2$.
- Collect data and compute the sample means and SEs.
  - Point estimate of the difference: $\bar{y} - \bar{x}$.
  - SE of the difference: $SE_{diff} = \sqrt{SE(\bar{x})^2 + SE(\bar{y})^2}$.
- Approximate 95% CI for $\mu_1 - \mu_2$ is:

$$\bar{y} - \bar{x} \pm 2SE_{diff}.$$

- When samples are small, then we need to be a bit more careful about the multiplier for the SE.

In the Glass Example...

- $\bar{x} = 5.3, S = 0.9, SE(\bar{x}) = 0.9/\sqrt{10} = .285.$
- $\bar{y} = 5.9, S = 0.85, SE(\bar{y}) = 0.85/\sqrt{9} = .283.$
- observed difference is $\bar{y} - \bar{x} = 0.6.$
- Standard error for this difference is $SE_{diff} = \sqrt{.285^2 + .283^2} = 0.402.$
- A 95% CI for the difference $\mu_1 - \mu_2$ is:

$$5.9 - 5.3 \pm 1.96 \times 0.402 = [-0.188,\ 1.388]\,.$$

- We conclude that plausible values for the difference are -1.188 to 1.388.

Hypothesis Tests and Confidence Intervals

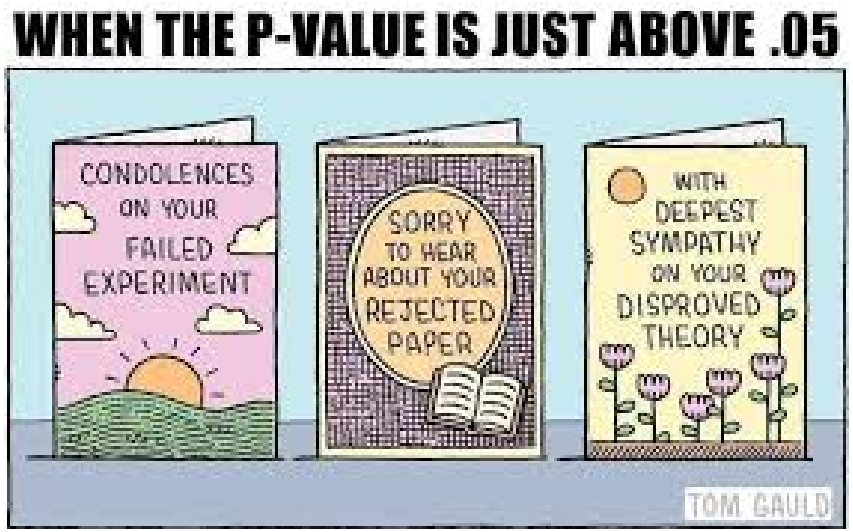
- There is a very close relationship between tests to compare two means and interval estimates.
- Confidence interval (CI) gives range of plausible values (e.g., for the difference in two means).
- Test evaluates whether a specific value (e.g., zero in the two-sample test) is a plausible value.
- Therefore, when the interval covers the value 0, the corresponding two-tailed test would have failed to reject H_0.
- In general, we fail to reject $H_0 : \mu_1 - \mu_2 = m$ at level α if the $100(1-\alpha)\%$ CI for the difference includes the value m.
- This relationship does not hold when the alternative hypothesis is one-sided, e.g., $\mu_1 - \mu_2 > m$.

Hypothesis Testing - Discussion

- Hypothesis testing does not treat the two hypotheses symmetrically (null is given priority).
  - This is appropriate if there is reason to prefer the null hypothesis until there is significant evidence against it.
  - We don't always want this to be the case (more on this later in forensic context).
- *p*-values depend heavily on the sample size:
  - If you have the same means and standard deviations and increase the sample size the result will be more significant.
- Interpretation can be tricky:
  - Rejecting the null hypothesis does not mean that one has found an important difference.
  - Important to consider the size of the observed difference.
  - Failing to reject the null hypothesis does not mean that the null hypothesis is true.
  - Important to consider the "power" of the test (how often would it reject if the alternative were true).

$p-$values in the Hot Seat

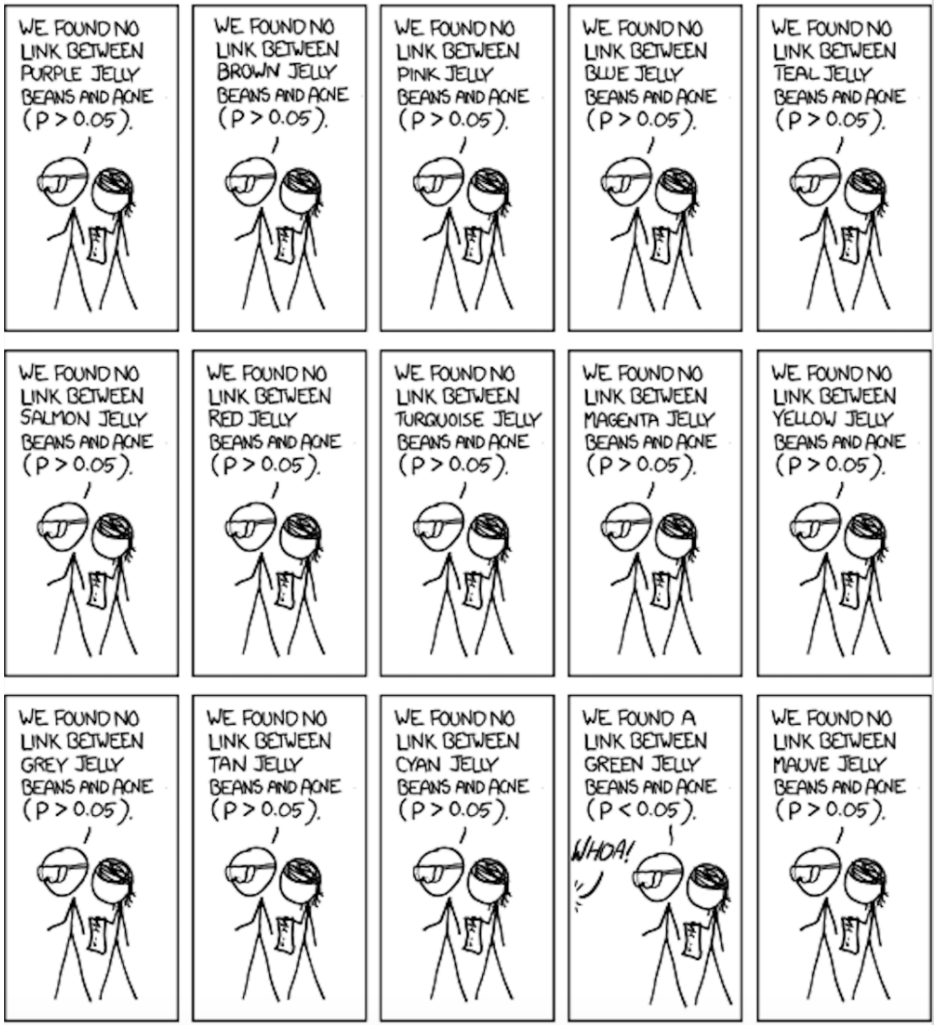
- $p-$values have been misinterpreted and misused.
- **A $p-$value of 0.05 does NOT mean that the null is true with probability 0.05**.
- The 0.05 threshold is arbitrary: $p-$value of 0.049 results in *statistically significant* result, but value of 0.051 does not.

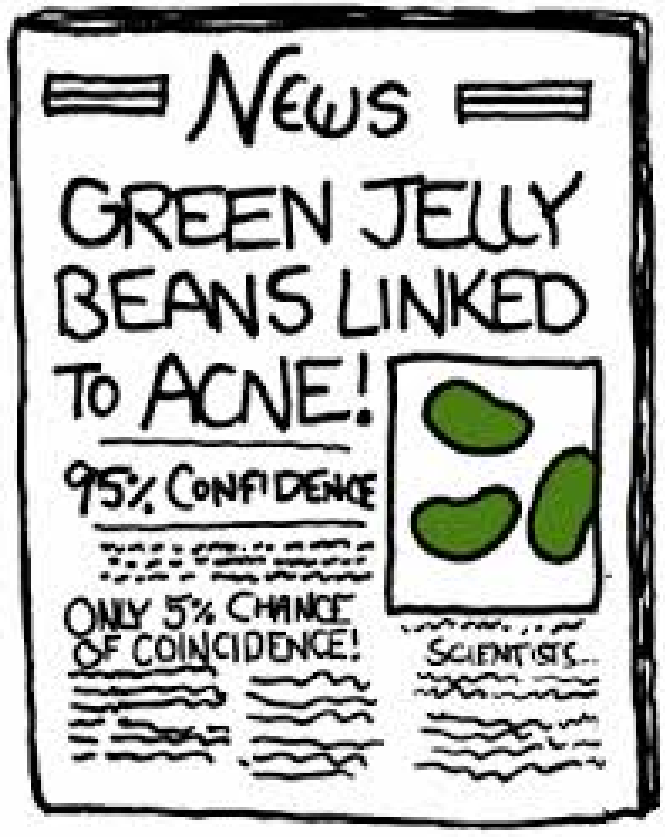


Beware of Multiple Tests

- Suppose you carry out two independent tests of hypotheses.
- In each, you fix your type I error to 0.05.
- Assume that H_a is true in both tests. Then there are four possible outcomes:
  - You correctly reject H_0 both times. Probability is $0.95^2 = 0.903$.
  - You incorrectly fail to reject H_0 both times. Probability is $0.05^2 = 0.0025$. Or you correctly reject H_0 in one test and incorrectly fail to reject in the other. Probability is $2 \times 0.95 \times 0.05 = 0.095$.
- Note that the only correct outcome is the first, and it has a confidence of only 0.9.
- When we carry out multiple tests, the *collective confidence level* is eroded rapidly.
- For 5 tests, confidence is about $0.95^5 = 0.77$; it is likely that at least one "significant" result is due to chance alone.

Chance Results





Practical Significance

- A very small $p-$value may indicate a statistically significant result that has **no practical significance**.
- It is better to report an *effect size*, which in the case of a test for difference of two means, is just the standardized difference:

$$d = \frac{\bar{x} - \bar{y}}{S_{pooled}},$$

where the pooled standard deviation is:

$$S_{pooled} = \sqrt{\frac{(n1-1)S_X^2 + (n2-1)S_Y^2}{n1+n2-2}}.$$

- The statistic d is known as Cohen's d.
- Values of d around 0.2 are low, around 0.5 are medium and around 0.8 or above are large.

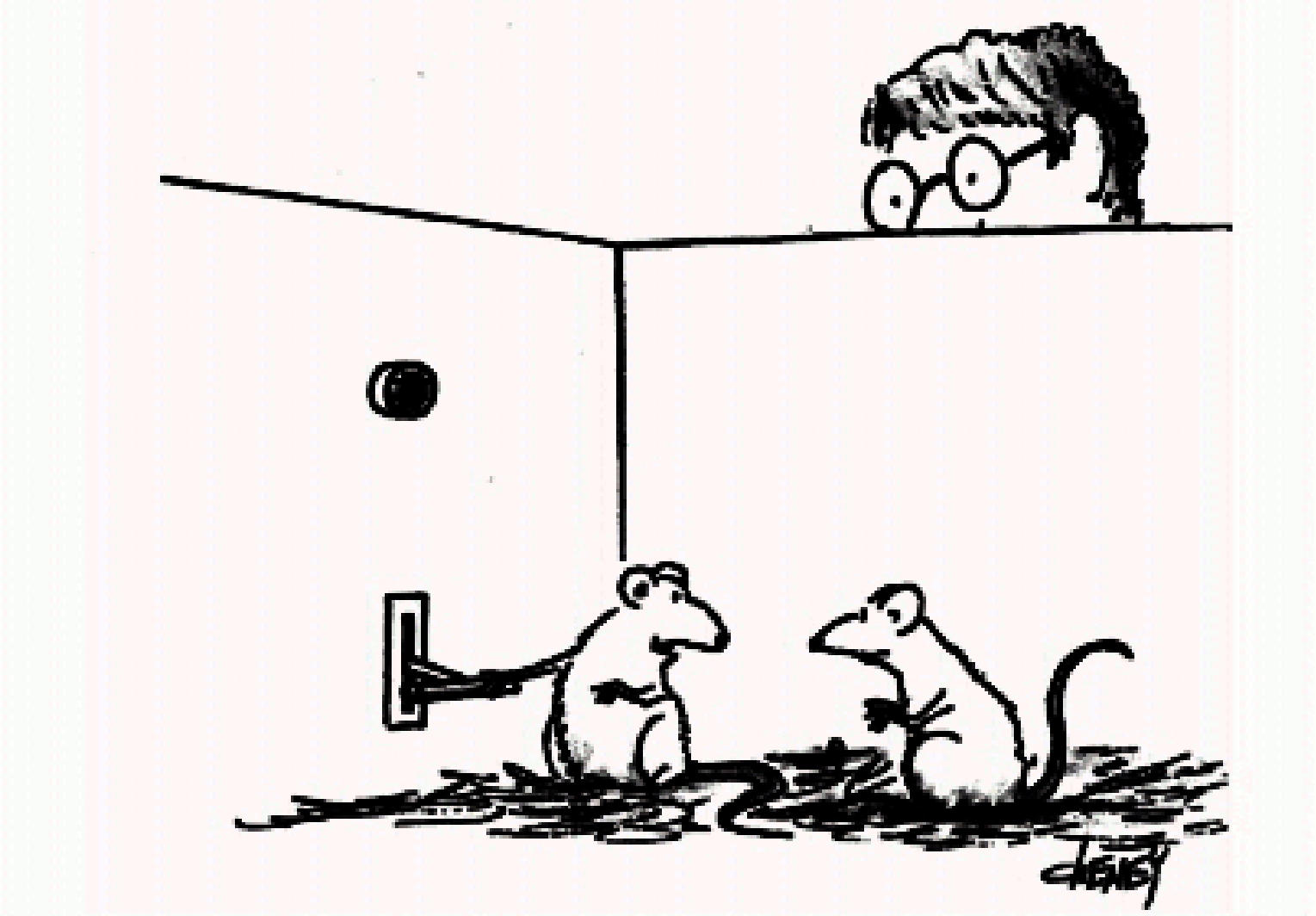
Approximate Sample Size

- How large a sample do I need to be able to detect an effect size of d with power $1-\beta = 0.8$ and confidence $1-\alpha = 0.95$?
- A rough estimate for the size of the sample from **each** population is:

$$n_i = \frac{2(Z_{\alpha/2} + Z_{1-\beta})^2}{d^2} = \frac{2(1.96 + 0.84)^2}{d^2} \approx \frac{16}{d^2}.$$

- 1.96 is the standard normal quantile at 0.025 and 0.84 is the standard normal quantile at 0.8.
- If $d = 0.5$, then $n_i = 64$, so total sample size is $n_1 + n_2 = 128$.
- If $1-\beta$ is 0.9 or 0.95, the corresponding $z-$values are 1.28 and 1.65.

Interpret Results Cautiously



It's a rather interesting phenomenon. Every time I press this lever, that post-graduate student breathes a sigh of relief.

A Limitation of Hypothesis Testing

- Suppose that you are asked to determine whether a new instrument produces measurements that are significantly more precise than the old one.
- The new instrument would cost the lab a LOT of money, so you do not want to incur the cost unless the difference in precision is substantial.
- Like a good statistician, you get n_1, n_2 measurements of reference items using both instruments and set up a test:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_a : \mu_1 \neq \mu_2.$$

- In this particular case, you might NOT wish to reject the null.
- In fact, I would like to show that the null is true.

A Limitation of Hypothesis Testing

- You carry out the test and fail to reject the null. Can we conclude that the precision of the instruments is the same?
- Not really...
  - I can always "rig" the test to fail to reject. Just set $n_1 = n_2$!
  - In traditional hypothesis testing, we *assume* that the null is true and then challenge that claim.
- What we require, is an approach that assumes that the means are different.
- We then collect evidence to challenge that assumption.

Equivalence Testing

- An alternative to hypothesis testing is *equivalence testing*.
- In equivalence testing, we turn hypothesis testing around and formulate the following hypotheses:

$$H_0 : |\mu_1 - \mu_2| > \Delta \text{ versus } H_a : |\mu_1 - \mu_2| \leq \Delta.$$

- The value Δ is the smallest difference between μ_1 and μ_2 that we consider to be of *practical significance*.
- If the difference between the means is within the equivalence range, we consider the means to be *equivalent*.
- Note that the difference between the means could be statistically significant but equivalent, or the other way around.
- Multiple approaches to equivalence testing. Most common one is the *two one-sided tests* or *TOST* approach.

Equivalence Testing

- Use δ to denote the difference between population means, so that $\delta = \mu_1 - \mu_2$.
- We set lower and upper equivalence bounds Δ_L and Δ_U, respectively, so that $\Delta = \Delta_U - \Delta_L$.
- In TOST, we test two null hypothesis against the correspond alternatives:

$$\begin{aligned} H_0 1: &\quad \delta \leq \Delta_L \\ H_0 2: &\quad \delta \geq \Delta_U. \end{aligned}$$

- We conclude that the means are *equivalent* at level α only if both null hypotheses are rejected.
- Since both tests are one-sided tests, we compute two critical values and two $p - values$ to decide.

Equivalence Testing - Calculations

- The computations are very similar to the ones we have already done.
- We compute two $t-$statistics:

$$t_L = \frac{\hat{\delta} - \Delta_L}{SE_{\hat{\delta}}} \text{ and } t_U = \frac{\hat{\delta} - \Delta_U}{SE_{\hat{\delta}}},$$

where $\hat{\delta} = \bar{x}_1 - \bar{x}_2$ and

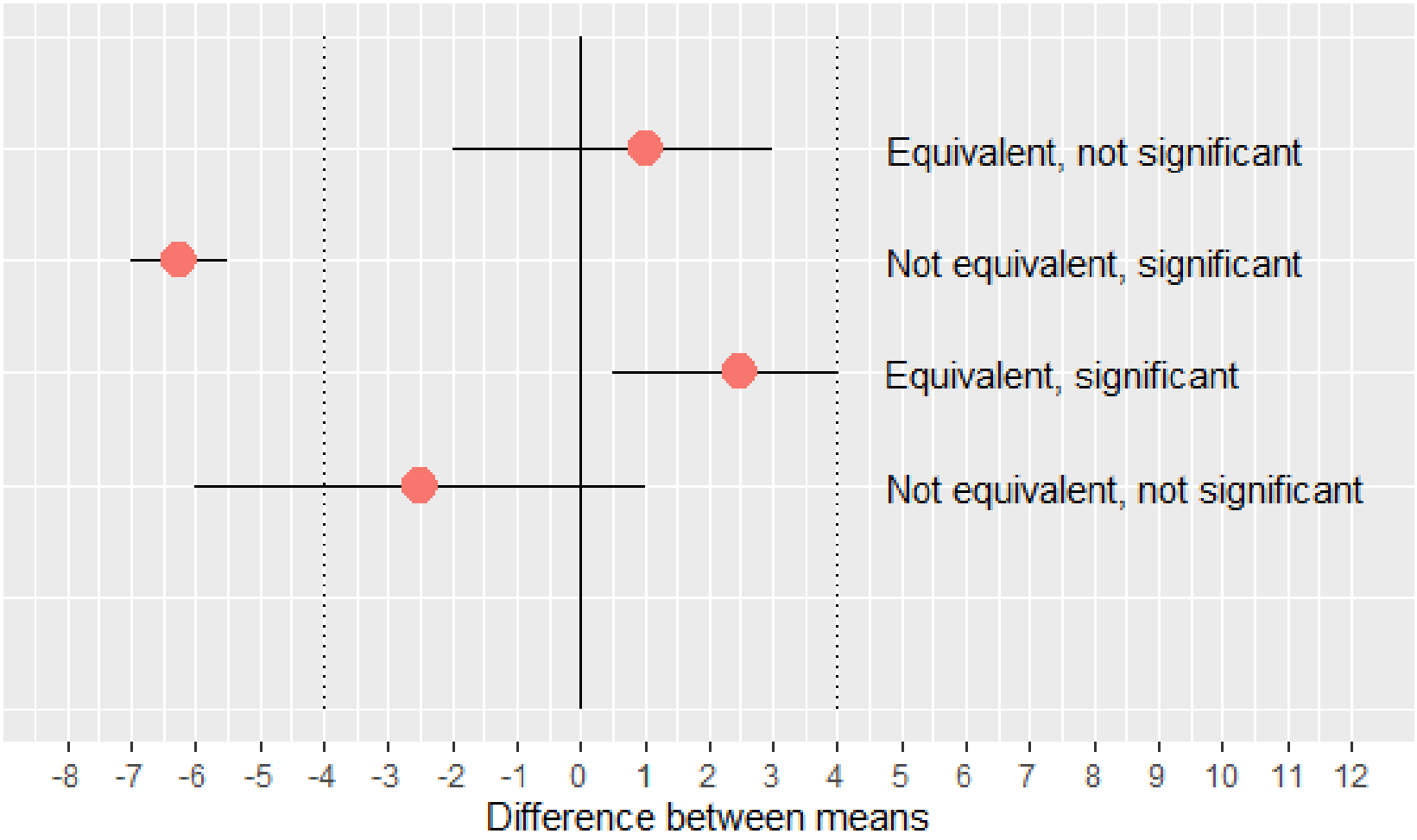
$$SE_{\hat{\delta}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

- At level α, we reject $H_0 1$ when $t_L \leq t_\alpha^*$ and reject $H_0 2$ when $t_U \geq t_{1-\alpha}^*$.

Equivalence Testing - Observations

- The challenging part in equivalence testing is to choose the equivalence bounds Δ_L, Δ_U.
- If we know the size of the effect that we consider practically important, we can set bounds in the units of the measurements.
- Else, we can standardize effect size and use Cohen's d values, so that a typical $\Delta_L = -0.3$ and a typical $\Delta_U = 0.3$.
- Instead of computing the $t-$statistics, the two null hypothesis can be tested using $100(1-2\alpha)\%$ confidence intervals for δ:
  - If the CI is fully within the equivalence bounds, we declare equivalence.
  - Otherwise, we declare non-equivalence.

Illustration



Statistical Thinking for Forensic Practitioners

Part 6: Inference (Video 5)

Dr. Alicia Carriquiry

Contents of Video 5

- Example of equivalence testing using TOST and CI approaches.
- Hypothesis testing for proportions and differences of proportions.
- Adjusting confidence levels when doing multiple tests of hypothesis.

Example using TOST

- The data we use for this example were adapted from Jenna A. Campbell's research paper published by Duquesne University in 2018.
- The question of interest is whether the concentration of Ba in GSR produced from firing 9mm Federal Premium ammunition is equivalent to the concentration measured in GSR produced from Remington 9mm ammunition.
- Experiment:
  - 9mm Beretta 92 FS pistol.
  - Five shots using each of the two types of ammo.
  - Extract GSR from hands of shooter.
  - Measure Ba (in % weight) using SEM.

Example using TOST

- Data:
  Federal: $\{$ 16.75 17.01 16.49 19.19 13.87 $\}$ Remington:
  $\{$ 13.75 12.08 14.42 13.72 12.20 $\}$.
- Means and SDs: $\bar{x} = 16.66$, $S_x = 1.90$, and
  $\bar{y} = 13.23$, $S_y = 1.04$.
- We decide that if the Ba difference in % wgt between the two brands of ammunition is within 0.5%, we declare equivalence.
- Equivalence limits: $\Delta_L = -0.5\%$, $\Delta_U = 0.5\%$.
- These limits are chosen by the investigator and correspond to values of the difference that are not considered to be of practical importance.
- Choose a confidence level of 0.95.

Example using TOST

- The two one-sided null hypotheses we test are:
  $H_{0_1} : \mu_1 - \mu_2 \leq \Delta_L$ and $H_{0_2} : \mu_1 - \mu_2 \geq \Delta_U$,
  where μ_1, μ_2 denote the mean Ba concentration of 9mm Federal ammunition and of 9mm Remington ammunition, respectively.
- Corresponding t-statistics:

$$t_L = \frac{\bar{x} - \bar{y} - \Delta_L}{SE_{\bar{x}-\bar{y}}}, \text{ and } t_U = \frac{\bar{x} - \bar{y} - \Delta_U}{SE_{\bar{x}-\bar{y}}},$$

where $SE_{\bar{x}-\bar{y}} = (S_x^2/n_1 + S_y^2/n_2)^{1/2}$.

Example using TOST

- With the sample statistics we have computed: $SE_{\bar{x}-\bar{y}} = (1.9^2/5 + 1.04^2/5)^{1/2} = 0.97$.
- Then:

$$t_L = \frac{16.66 - 13.23 - (-0.5)}{0.97} = \frac{3.93}{0.97} = 4.05,$$

$$t_U = \frac{16.66 - 13.23 - 0.5}{0.97} = \frac{2.93}{0.97} = 3.02.$$

- We reject both H_{0_1} and H_{0_2} at level α, since the critical value is $t_L^* = t_{0.05,8} = -1.86$ and $t_U^* = 1.86$.
- A $100(1-2\alpha)$CI for $\mu_1 - \mu_2$ is

$$\bar{x} - \bar{y} \pm t_{0.05,8} SE_{\bar{x}-\bar{y}} = 3.43 \pm 1.86 \times 0.97 = [1.66,\ 5.23]\,.$$

- Since the interval is fully outside the equivalence region, we reach the same conclusion as above.

Hypothesis test for a proportion

- Suppose we wish to test whether a population proportion π is equal to some value π_0.
- The steps are the same as before:
  - Set up the null and alternative hypotheses.
  - Choose a confidence level $1 - \alpha$.
  - Draw sample from population and compute sample statistics and test statistic.
  - Decide, comparing the test statistic to the critical value or using a p-value and comparing against α.
- As before, we can test the null against a two-sided or a one-sided alternative.
- Recall that an estimator for π is the sample proportion $\hat{p}$ and an estimator for the SE of $\hat{p}$ is $[\pi(1-\pi)/n]^{1/2}$.

Hypothesis test for a proportion

- National sales data for 2017 indicate that 21.1% of all men's athletic shoes purchased in the US were made by Nike.
- But we wonder whether there may be some regional variation, and are interested in whether Nike has the same market share in Iowa.
- We formulate a two-sided test of hypothesis:

$$H_0 : \pi = \pi_0, \text{ versus } H_a : \pi \neq \pi_0,$$

  where $\pi_0 = 0.211$.
- Choose an $\alpha = 0.05$ so the confidence level is 0.95.

Hypothesis test for a proportion

- We interview a random, representative sample of size 200 individuals aged 12 to 80 years, all men, and ask:
  - How many pairs of shoes purchased in 2017, 2018 and 2019.
  - How many of those were Nike.
- The total number of pairs of shoes purchased was $N = 760$, and of those, $Y = 190$ were Nike.
- Estimate of π is $\hat{p} = 190/760 = 0.25$.
- The SE of $\hat{p}$ depends on π, so what should we use to compute it?
- Since the null hypothesis establishes that $\pi = \pi_0 = 0.211$, then this is the value we plug in, to get:

$$SE_{\hat{p}} = \sqrt{\frac{0.211(1 - 0.211)}{200}} = 0.029.$$

Hypothesis test for a proportion

- The test statistic here is a $z-$statistic because under the H_0, the SE of $\hat{p}$ is known.
- The statistic is:

$$z = \frac{0.25 - 0.211}{0.029} = \frac{0.039}{0.029} = 1.34.$$

- The critical value is $z_{\alpha/2} = 1.96$. Since 1.34 is not in the critical region, we fail to reject H_0.
- Alternatively, the $p-$value is the probability of observing a value more extreme than 1.34:

$$p - \text{value} = 2\text{Prob}(z \geq 1.34) = 2 \times 0.09 = 0.18,$$

which also leads to failing to reject the null.

- Conclusion: There is no evidence to suggest that the market share of Nike shoes in Iowa is different from its national market share.

Testing using the CI approach

- Since the test of hypothesis was two-sided, we could have also used a $100(1-\alpha)\%$ CI interval to decide between H_0 and H_a.
- A 95% CI for π is:

$$\begin{aligned} \hat{p} \pm 1.96 \times \left[\frac{\pi_0(1-\pi_0)}{n}\right]^{\frac{1}{2}} &= 0.25 \pm 1.96 \times 0.029 \\ &= 0.25 \pm 0.057 \\ &= [0.193 \text{ to } 0.307]\,. \end{aligned}$$

- Since the interval covers the hypothesized value 0.211, we cannot reject H_0, and conclude that 0.211 appears to be a plausible value for the Nike market share in Iowa.

Testing for difference in proportions

- Suppose that now we have two populations and would like to test for the difference in the proportion of some attribute.
- A two-tailed test of hypothesis is:
  $H_0 : \pi_1 = \pi_2$ versus $H_a : \pi_1 \neq \pi_2$.
- We obtain a sample of size n_1 from one population and of size n_2 from the other.
- We observe Y_1 and Y_2 successes, respectively.
- If H_0 is true, then the two proportions are equal to the same value and I could estimate that value pooling the data from both samples.

Testing for difference in proportions

- To compute the SE of the difference, we reason as follows: If H_0 is true, then the two proportions are equal to the same value and I could estimate that value pooling the data from both samples.
- Estimator of common population proportion:

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2},$$

with standard error:

$$\begin{aligned} SE_{\hat{p}} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} \\ &= \sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

Testing for difference in proportions

- The $z-$statistic is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (\pi_1 - \pi_2)}{SE_{\hat{p}}} = \frac{(\hat{p}_1 - \hat{p}_2)}{SE_{\hat{p}}},$$

  where the term $(\pi_1 - \pi_2)$ disappears because under H_0, the two population proportions are equal.
- For confidence level $1 - \alpha$, reject H_0 if $|z| \geq z_{\alpha/2}$.
- For $\alpha = 0.05$, we reject H_0 if $z \leq -1.96$ or $z \geq 1.96$.

Difference in proportions - Example

- We have estimated the Nike market share in Iowa for men's shoes, and now would like to know whether the Nike market share in Texas is similar.
- We interview a representative random sample of size 300 of TX males aged 12-80.
- In the same period, the TX men purchased an average of 3.7 pairs of shoes each (so $n_T = 1110$) of which $Y_T = 212$ were Nike.
- Now $\hat{p} = (190 + 212)/(760 + 1110) = 0.215$, and the SE is

$$SE_{\hat{p}} = \sqrt{0.215(1 - 0.215)}\sqrt{\frac{1}{760} + \frac{1}{1110}} = 0.412 \times 0.0471 = 0.019.$$

Difference in proportions - Example

- The test statistic is:

$$z = \frac{0.25 - 0.19}{0.019} = 3.158.$$

- For $\alpha = 0.05$ the statistic falls inside the critical region because $3.158 > 1.96$.
- The $p-$value of this test is $2\mathrm{P}(z > 3.158)$ which is $2 \times 0.000794 = 0.002$.
- We reject the null hypothesis and conclude that evidence supports the proposition that the Nike market shares in Iowa and in Texas are different.
- Because the $p-$value was $\leq$ than 0.01, we would have concluded H_a even with 99% confidence.

Adjusting confidence levels

- We mentioned that the confidence level for a family of tests decreases as the number of tests increases.
- Suppose that in addition to testing hypotheses about the difference in Nike market share in Iowa and Texas, we also tested the same hypotheses for Adidas, Puma, New Balance and Under Armour.
- The total number of tests is now 5 and if we start with a $(1-\alpha)$ confidence level for each test, an approximate confidence for all five tests is only $(1-\alpha)^5$.
- For $\alpha = 0.1$, the overall confidence is approximately 0.59. For $\alpha = 0.05$, we get 0.77.

Bonferroni correction

- Simplest way to correct type I error probabilities with multiple testing.
- Suppose I carry out m tests of hypotheses.
- The Bonferroni correction is to use α/m in each test instead of the original α.
- For $\alpha = 0.05$ and $m = 5$, the Bonferroni adjusted test-wise type I error is $\alpha_B = 0.05/5 = 0.01$.
- If the five tests are independent, then the family-wise confidence level is

$$(1 - \alpha_B)^5 = 0.99^5 = 0.951.$$

- By "independent", we mean that whether we carry out a test or not does not depend on the results of the other tests.

Bonferroni correction

- Shortcoming of Bonferroni is that when m is very large, $\alpha_B \longrightarrow 0$, so we end up failing to reject all the H_0.
- In this case, we would have no *false positives* (incorrect rejections of H_0) but we would have a lot of *false negatives* (failing to reject a false H_0).
- We say that the Bonferroni correction is *conservative*.

Controlling a false discovery rate (FDR)

- An alternative is to fix the *false discovery rate*.
- The FDR is the proportion of false positives among all the tests where we rejected H_0.
- Suppose that I carry out many tests and reject H_0 in R of those tests.
- Some of those positive decisions (H_a decisions) will be correct, some will be *false positives* or *false discoveries*.
- We use T to denote true positives and F to denote false positives, so that $R = F + T$.

False discovery rate

- The FDR is defined as

$$FDR = \frac{F}{F + T}.$$

- The idea is to choose an acceptable FDR and then adjust the original $p-$values so that the proportion of false positives does not exceed the FDR.

Controlling a false discovery rate (FDR)

- The most popular method to control FDR was proposed by Benjamini and Hochberg (1995).
- Steps are the following:
  - Compute $p-$values for your m tests. Call them $p_1, p_2, ..., p_m$ and order them from smallest to largest.
  - The ordered $p-$values are denoted $p_{(1)}, p_{(2)}, ..., p_{(m)}$.
  - To control FDR at level α, select the largest k such that $p_{(k)} < \alpha k/m$.
  - Reject the null hypothesis $H_{0_1}, H_{0_2}, ..., H_{0_k}$.
- It is much easier to look at an example.

BH for controlling FDR

- Suppose that we carry out $m = 10$ hypothesis tests.
- Suppose further that we reject H_0 in six of those tests.
- The six $p-$values we got are: 0.04, 0.002, 0.045, 0.008, 0.039 and 0.047.
- We order them from smallest to largest, and compare each to $k\alpha/m$.

$p_{(i)}$:	0.002	0.008	0.039	0.04	0.045	0.047
k:	1	2	3	4	5	6
$k\alpha/m$:	0.005	0.01	0.015	0.02	0.025	0.03

- We would end up concluding H_a only for tests 2 and 4, in the original order shown in bullet 3 above.

BH for controlling FDR

- We can plot k (x-axis) against $k\alpha/m$ (y-axis) using blue dots.
- Ordered $p - values$ are red dots.
- Reject H_0 when red dots are below line.