

CSCI 3022-002 Intro to Data Science

Small Samples

- ▶ NO Short HW for next week.
- ▶ NB day Friday.

p-values

Definition: p-value:

A *p-value* is the probability, under the null hypothesis, that we would get a test statistic at least as extreme as the one we calculated.

$$\begin{aligned}
 & \left\{ \begin{array}{l} \text{as large as} \\ \text{as small as} \\ \text{as absolute value large as} \end{array} \right. \begin{array}{l} \text{---} \\ \text{---} \\ \text{---} \end{array} \begin{array}{l} H_0: \mu > \mu_0 \\ H_0: \mu < \mu_0 \\ H_0: \mu = \mu_0 \end{array} \begin{array}{l} ; \text{ reject if } \bar{X} \text{ large} \\ ; \text{ " " } \bar{X} \text{ small} \\ ; \text{ reject if } |\bar{X}| \text{ large.} \end{array}
 \end{aligned}$$

$$p(Z > z_{\alpha/2})$$

$$P(\text{a new } \bar{X} > \text{our } \bar{X}) = p.$$

Idea: So, the smaller the p-value, the more evidence there is in the sample data against the null hypothesis.

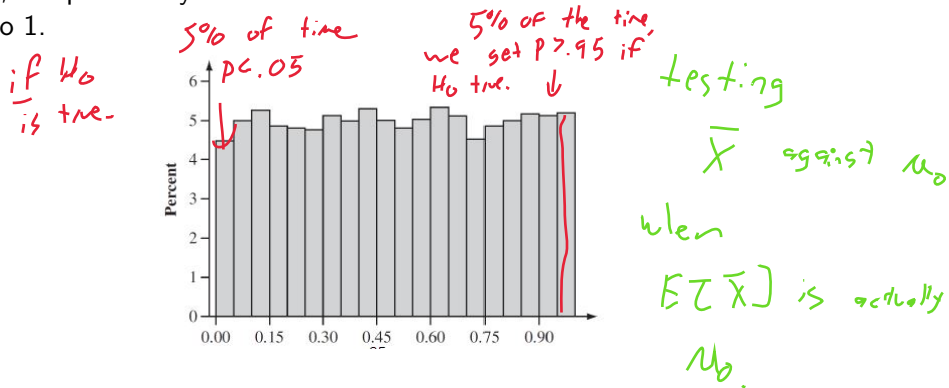
1. This probability is calculated assuming that the null hypothesis is true.
2. Beware: The p-value is not the probability that H_0 is true, nor is it an error probability!
3. The p-value is between 0 and 1.

Compare α to p-value.

Distribution of p-values

Figure below shows a histogram of the 10,000 P-values from a simulation experiment under a null $\mu = 20$ (with $n = 4$ and $\sigma = 2$).

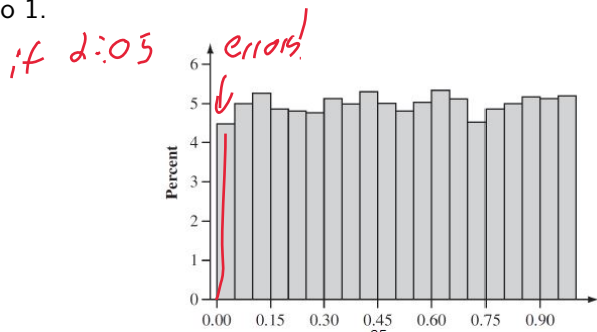
When H_0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.



Distribution of p-values

Figure below shows a histogram of the 10,000 P-values from a simulation experiment under a null $\mu = 20$ (with $n = 4$ and $k = 2$).

When H_0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.



These data comes from a process where the null hypothesis is *TRUE*. Rejecting the null hypothesis would be an error.

Distribution of p-values

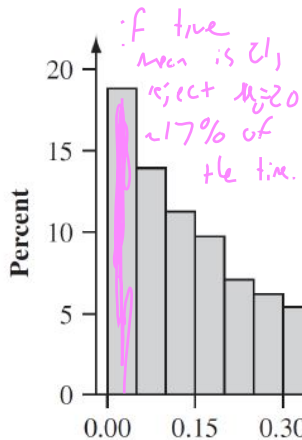
About 4.5% of these p-values are in the interval from 0 to .05.

Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests.

If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run, 5% of the p-values would be in the first class interval.

Distribution of p-values

A histogram of the p-values when we simulate under an alternative hypothesis. There is a much greater tendency for the p-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$



testing $\mu_0 = 20$

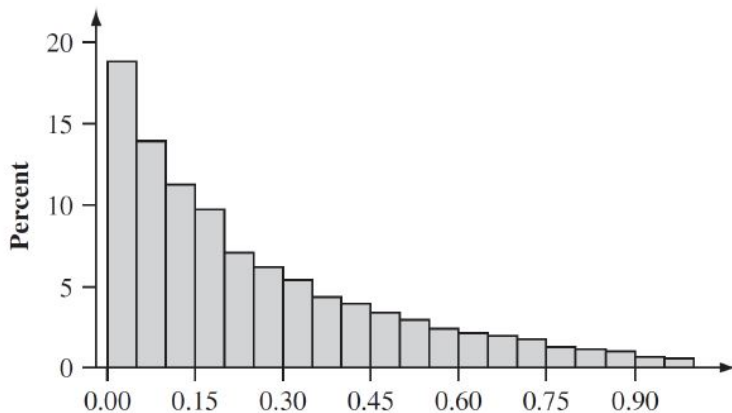
reality: $\mu = 21$

then \bar{X} tends to be close to 21

more often than 5%,
 $\bar{X} > 20$ by enough
 for us to reject.

Distribution of p-values

A histogram of the p-values when we simulate under an alternative hypothesis. There is a much greater tendency for the p-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$



Distribution of p-values

Again, H_0 is rejected at significance level .05 whenever the p-value is at most .05 (in the first bin).

Unfortunately, this is the case for only about 19% of the p-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed.

The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis (the “effect size” is small).

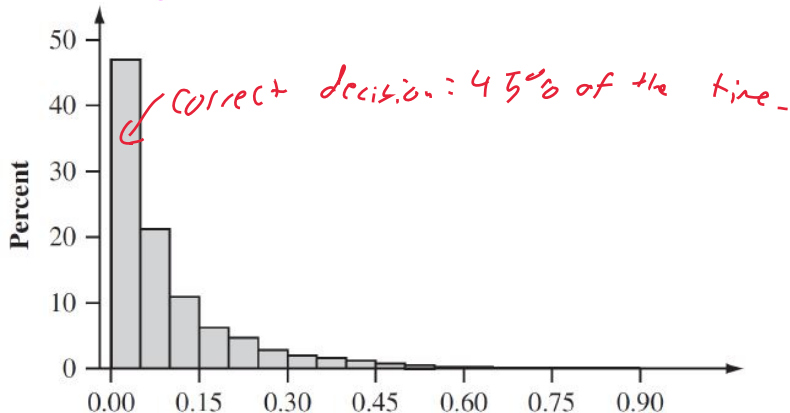
Distribution of p-values

The figure below illustrates what happens to the p-value when H_0 is false because

$$\mu = 22.$$

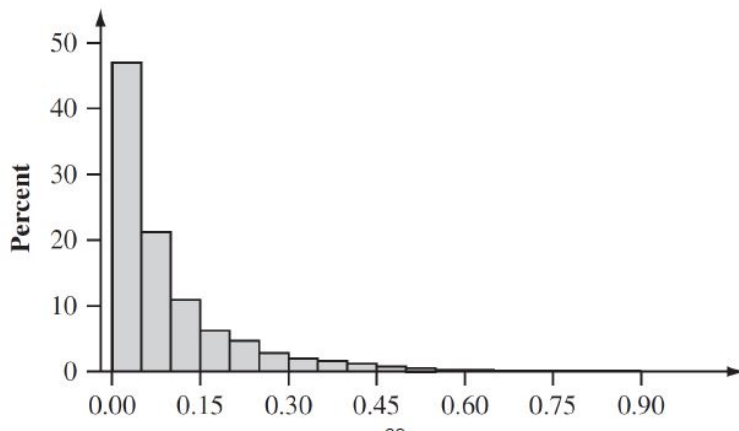
$$H_0: \mu_0 = 20$$

the
value



Distribution of p-values

The figure below illustrates what happens to the pvalue when H_0 is false because $\mu = 22$.



Again, these data come from (H_a) , but we reach the correct conclusion more often. The

Distribution of p-values

The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 21$.

In general, as μ moves further to the right of the null value 20, the distribution of the p-value will become more and more concentrated on values close to 0.

We are correct more if $|\mu - \mu_0|$ large.

Even here a bit fewer than 50% of the p-values are smaller than .05. So it is still slightly more likely than not that the null hypothesis is incorrectly not rejected. Only for values of much larger than 20 (e.g., at least 24 or 25) is it highly likely that the p-value will be smaller than .05 and thus give the correct conclusion.

Comparing 2 Means

Normal Populations with known variances:

If both populations are normal and independent, $\bar{X} - \bar{Y}$ is normally distributed with expected value $\mu_1 - \mu_2$ and standard deviation: $\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$. So:

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

better as
samples grow

Standardizing our estimator gives:

proposed
 $\Delta_0: \mu_1 - \mu_2$

on
average, correct estimate

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

Therefore, the $(1 - \alpha) \cdot 100\%$ confidence interval is:

test: $(\bar{X} - \bar{Y}) - \Delta_0$

$$Z = \frac{(\bar{X} - \bar{Y}) - \Delta_0}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$$

$$(\bar{X} - \bar{Y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

$A - B = 0$
 $\Leftrightarrow A = B$

Large Sample Tests

Example:

Data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds more than 200 calories per day the true average intake for those who don't typically eat fast food? Let's investigate by carrying out a test of hypotheses at a significance level of 0.05.

Large Sample Tests

Example:

Data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds more than 200 calories per day the true average intake for those who don't typically eat fast food? Let's investigate by carrying out a test of hypotheses at a significance level of 0.05.

$$H_0 : \mu_2 - \mu_1 = 200; \quad H_a : \mu_2 - \mu_1 > 200$$

Plan: Reject if $p < .05$.

Large Sample Tests

Example:

Data on daily calorie intake both for a sample of teens who said they did not typically eat fast food and another sample of teens who said they did usually eat fast food.

Eat Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide strong evidence for concluding that true average calorie intake for teens who typically eat fast food exceeds more than 200 calories per day the true average intake for those who don't typically eat fast food? Let's investigate by carrying out a test of hypotheses at a significance level of 0.05.

$$z_{stat} = \frac{(2637 - 2258) - (200)}{\sqrt{\frac{1519^2}{663} + \frac{1138^2}{413}}} = 2.20 \quad P(Z < 2.20) = .987$$

Handwritten notes: A red circle around (200) with an arrow pointing to it from the text "Δ = 200" above. A red checkmark to the right of the circle.

Comparing 2 Means: Small Sample

For large samples, the CLT allows us to use these methods we have discussed even when the two populations of interest are not normal.

In practice, it can happen that at least one sample size is small and the population variances have unknown values.

⚠ if each sample is normal

Without the CLT at our disposal, we proceed by making specific assumptions about the underlying population distributions.

Comparing 2 Means: Small Sample

When the population distributions are both normal, the standardized variable

has approximately a t distribution with df ν estimated from the data by:

Comparing 2 Means: Small Sample

When the population distributions are both normal, the standardized variable

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \quad \text{C now with } s, \text{ not } \sigma!$$

has approximately a t distribution with df ν estimated from the data by:

recall

1 Sample testing $\nu = n - 1$

2 Samples • lazy way: $\nu = \min \begin{cases} n-1 \\ m-1 \end{cases}$

Comparing 2 Means: Small Sample

When the population distributions are both normal, the standardized variable

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

has approximately a t distribution with df ν estimated from the data by:

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n} \right)^2}{\frac{(s_1^2/m)^2}{m-1} + \frac{(s_2^2/n)^2}{n-1}}$$

DO
NOT
MEMORIZE

Stats. w. m. interval
Stats. f. interval

Comparing 2 Means: Small Sample

The two-sample t confidence interval for $\mu_1 - \mu_2$ with confidence level $(1 - \alpha) \cdot 100\%$ is then:

$$(\bar{X} - \bar{Y}) \pm t_{\alpha/2, \nu} \sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$$

t
 versus $z_{\alpha/2}$

Test Procedures for Normal Populations with Unknown Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

Alt Hypothesis

Rejection Region

p-value:

Test Procedures for Normal Populations with Unknown Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

$$\Delta_0 = 5$$

is $\mu_1 > \mu_2 + 5$

testing "is method 1

Test statistic value:

$$t_{stat} = \frac{(\bar{X} - \bar{Y}) - (\Delta_0)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

5 points >

than method 2"

Alt Hypothesis

Rejection Region

p-value:

Test Procedures for Normal Populations with Unknown Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

$$t_{stat} = \frac{(\bar{X} - \bar{Y}) - (\Delta_0)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Alt Hypothesis

Rejection Region

p-value:

$H_a : \mu > \mu_0$

$H_a : \mu < \mu_0$

$H_a : \mu \neq \mu_0$

Test Procedures for Normal Populations with Unknown Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

$$t_{stat} = \frac{(\bar{X} - \bar{Y}) - (\Delta_0)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

<u>Alt Hypothesis</u>	<u>Rejection Region</u>	<u>p-value:</u>
$H_a : \mu > \mu_0$	$t_{stat} > t_{\alpha, \nu}$	
$H_a : \mu < \mu_0$	$t_{stat} < -t_{\alpha, \nu}$	
$H_a : \mu \neq \mu_0$	$ t_{stat} > t_{\alpha/2, \nu}$	

Test Procedures for Normal Populations with Unknown Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

$$t_{stat} = \frac{(\bar{X} - \bar{Y}) - (\Delta_0)}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \mu \geq \mu_0$	$t_{stat} \geq t_{\alpha, \nu}$	$P(T > t_{stat})$
$H_a : \mu \leq \mu_0$	$t_{stat} \leq -t_{\alpha, \nu}$	$P(T < t_{stat})$
$H_a : \mu \neq \mu_0$	$ t_{stat} > t_{\alpha/2, \nu}$	$P(T > t_{stat})$

↑
2-sided naturally!

Test for Equivalence of Proportions

Theoretically, we know that:

Recall: $Z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

1-sample

has approximately a standard normal distribution.

When $H_0 : p_1 - p_2 = 0$ is true, we have $p_1 = p_2$, which simplifies this:

However, this Z cannot serve as a test statistic because the value of p is unknown; H_0 asserts only that there is a common value of p , but does not say what that value is.

Test for Equivalence of Proportions

Theoretically, we know that:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

$p_i \rightarrow \hat{p}_i$

has approximately a standard normal distribution.

if we know

p_1, p_2, \dots

When $H_0 : p_1 - p_2 = 0$ is true, we have $p_1 = p_2$, which simplifies this:

hypothesis

However, this Z cannot serve as a test statistic because the value of p is unknown; H_0 asserts only that there is a common value of p , but does not say what that value is.

Test for Equivalence of Proportions

Theoretically, we know that:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

has approximately a standard normal distribution.

When $H_0 : p_1 - p_2 = 0$ is true, we have $p_1 = p_2$, which simplifies this:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}}$$

However, this Z cannot serve as a test statistic because the value of p is unknown; H_0 asserts only that there is a common value of p , but does not say what that value is.

Test for Equivalence of Proportions

Null: $p_1 = p_2$ if this is true, all data have the same p .

Under the null hypothesis, we assume that $p_1 = p_2 = p$, instead of separate samples of size m and n from two different populations (two different binomial distributions).

So, we really have a single sample of size $m + n$ from one population with proportion p .

The total number of individuals in this combined sample having the characteristic of interest is $X + Y$.

The estimator of p is then:

$$\hat{p}_x = \frac{X}{n} \quad \hat{p}_y = \frac{Y}{m}$$

$$\text{all } \hat{p} = \frac{X+Y}{n+m} \quad \text{combining data}$$

Test for Equivalence of Proportions

Under the null hypothesis, we assume that $p_1 = p_2 = p$, instead of separate samples of size m and n from two different populations (two different binomial distributions).

So, we really have a single sample of size $m + n$ from one population with proportion p .

The total number of individuals in this combined sample having the characteristic of interest is $X + Y$.

The estimator of p is then: $\hat{p} = \frac{X+Y}{n+m}$

Test for Equivalence of Proportions

Using \hat{p} and $1 - \hat{p}$ in place of p and $1 - p$ in our old equation gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

Alt Hypothesis

Rejection Region

p-value:

Test for Equivalence of Proportions

Using \hat{p} ; $1 - \hat{p}$ in place of p and $1 - p$ in our old equation gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

diff in means versus proposed

$$z_{stat} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

if H_0 true,

Alt Hypothesis

Rejection Region

p-value:

$$p_1 = p_2 \approx \hat{p} = \frac{x+y}{n+m}$$

$$= \frac{\# \text{ successes}}{\# \text{ total}}$$

Test for Equivalence of Proportions

Using and \hat{p} ; $1 - \hat{p}$ in place of p and $1 - p$ in our old equation gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

$$z_{stat} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

Alt Hypothesis

Rejection Region

p-value:

$H_a : \mu > \mu_0$

$H_a : \mu < \mu_0$

$H_a : \mu \neq \mu_0$

Test for Equivalence of Proportions

Using and \hat{p} ; $1 - \hat{p}$ in place of p and $1 - p$ in our old equation gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

$$z_{stat} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

<u>Alt Hypothesis</u>	<u>Rejection Region</u>	<u>p-value:</u>
$H_a : \mu > \mu_0$	$z_{stat} > z_\alpha$	
$H_a : \mu < \mu_0$	$z_{stat} < -z_\alpha$	
$H_a : \mu \neq \mu_0$	$ z_{stat} > z_{\alpha/2}$	



Test for Equivalence of Proportions

Using and \hat{p} ; $1 - \hat{p}$ in place of p and $1 - p$ in our old equation gives a test statistic having approximately a standard normal distribution when H_0 is true.

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often 0)

Test statistic value:

$$z_{stat} = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \mu > \mu_0$	$z_{stat} > z_\alpha$	$P(Z > z_{stat})$
$H_a : \mu < \mu_0$	$z_{stat} < -z_\alpha$	$P(Z < z_{stat})$
$H_a : \mu \neq \mu_0$	$ z_{stat} > z_{\alpha/2}$	$P(Z > z_{stat})$

We've looked at the following test statistics for hypothesis testing.

1. To compare proportions against a baseline or against each other, we use Z-statistics.

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad \text{OR} \quad \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

1-sample prop

2-sample

2. To compare means when the samples are large **or** underlying normal with known variances, we also use Z-statistics.

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

OR

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

large

OR

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

OR

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

large
x & y

3. To compare means when the samples are small **and** underlying normal, we use t-statistics.

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \quad \text{OR} \quad \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

Now what? if we draw a 90% CI of μ : $[a, b]$
 and ask is " c " inside that interval SAME as $H_0: \mu = c$.

We're almost done talking about CI's and Hypothesis tests. Where are our gaps? $\alpha = .10$.

1. We can compare samples or do inference on 1-2 samples when one of the following conditions is met:
 - 1.1 The sample or samples are $n > 30$ (or success/fail ≥ 10) (use Z !)
 - 1.2 The sample or samples are small and underlying normal (use t !)
2. What are we missing?
 - 2.1 The samples are small and not normal
 - 2.2 We aren't trying to do inference on means at all, but something else!

Bootstrapping is a catch-all to create *approximate* confidence intervals for any underlying population characteristic that we might care about.

To date, every one of our methods for confidence intervals and hypothesis testing have been based on the tests regarding the *mean*. We might want to test variances! Or medians! Or 87th percentiles!

We also might want to test means on small samples from *non-normal* populations. Data is often very expensive, in either time or money. Examples:

1. Data collected by aircraft
2. Polling data, which requires one-to-one human interactions
3. Seasonal ecological data, which may occur only once per calendar year

The **Bootstrap Principle** is a technique for *both* the “want more data” and “need other statistics” problems.

Bootstrapping

The **Bootstrap Principle** is a technique for *both* the “want more data” and “need other statistics” problems.

Definition: A *bootstrapped sample* is a set of n draws from the original sample set with replacement.

Example: Suppose we have the data set $X = [2, 2, 4, 7, 9]$. Some resamples might be:

1. $X_1 = [2, 4, 4, 4, 7]$

2. $X_2 = [4, 4, 4, 4, 4]$

3. $X_3 = [4, 2, 7, 9, 9]$

\hookrightarrow $\text{np.random.choice}([2, 2, 4, 7, 9])$

...each of those have their very own *sample statistics*!

Bootstrapping

A *bootstrapped sample* is a set of n draws from the original sample set with replacement.

As a rule-of-thumb, each bootstrapped sample should be of the same size as the original sample.

Proposition: A suitable estimate for the 95% confidence interval for the mean of the population of X is given by $[L, U]$, where L and U are the 2.5th and 97.5th sample percentiles of the set of means of a large number of bootstrapped resamples.

1000 new X 's. Sort them. Throw out lowest 2.5% & highest 2.5%.

Idea: Bootstrapping gives us a set of new X 's and new \bar{X} 's. The "middle 95%" of the bootstrapped \bar{X} 's should be in around the same place as the 95% CI for \bar{X} , regardless of distribution of individual X -values.

Bootstrapping solves all

$$\bar{x} = \sim 2500$$

$$[1, 2, 3, 10000]$$

same x 's: $[2, 1000, 3, 10000]$
 $: [1000, 1000, 1000, 1000]$

Bootstrapping for a CI around the mean is convenient, particularly when there are not enough samples to invoke the Central Limit Theorem.

Crucially, we can use the exact same procedure to estimate things besides means!

1. Medians — sample median
2. Standard Deviations — sample std.
3. Other measures that we may not even have theories for!

↳ sample 83rd percentile

Bootstrapping a median

Suppose we want a 90% CI for the variance of a data set. Code to **bootstrap**:

large or small (ish).

Bootstrapping a median

median
→ means

Suppose we want a 90% CI for the variance of a data set. Code to **bootstrap**:

1. `vars=[]`
`nsamp=10000` } initialize
2. `for i in range(nsamp):`
 `newX=np.random.choice(X, size=len(X), replace=True)`
 `vars.append(np.var(newX, ddof=1))` } make 10000 new samples
3. `CI= np.percentile(vars, [5,95])` } compute variance for each

list of 10000 variances

compute variance for each

middle 90% of sorted list of variances

Bootstrapping in general

This process: simulating a data set, calculating a desired *sample statistic* from it, and then creating a *distribution* of that sample statistic is called a *non-parametric* bootstrap since it doesn't make distributional assumptions.

Definition: *parametric* statistics assume that sample data comes from a population that follows a probability distribution on a fixed set of parameters.

Examples:

1. μ and σ are the parameters of the Normal distribution.
2. λ is the parameters of the Poisson and Exponential distributions.
3. p is the parameter of the geometric and Bernoulli distributions.

Parametric Bootstrapping

Sometimes we really want to know about various statistics on e.g. the Poisson or Exponential *without* solving some challenging integral or sum or whatever else equations.

Definition: *parametric* bootstraps estimate a CI for a desired property in two steps.

1. Estimate the parameters of the known distribution from your sample.
2. Draw bootstrap resamples from the distribution, *assuming* the estimated parameter
3. Compute a CI for the desired property from your resamples.

Parametric Bootstrapping

Example: If we want to estimate the median of a sample that we *assume* is Poisson, we might:

1. Assume the data is $\text{Pois}(\lambda)$. Estimate the parameter, e.g. $\lambda \approx \bar{X}$.
2. Simulate a bootstrapped sample from $\text{Pois}(\bar{X})$.
3. Create a CI for the median from that pool of bootstrapped samples.

Why make *more* assumptions, like assuming the distribution of the random variable at all? The advantage of the parametric bootstrap is that it can be shown to do a better job in particular scenarios.

The downside? The parametric bootstrap does a very poor job if the population does not have the same population as you assumed. This is called *model misspecification*, and is a risk **any** time we assume things have **any** underlying distribution, including in hypothesis testing!

Daily Recap

Today we learned

1. Comparing multiple samples for equivalence of the mean! Bootstrapping to skip all of that hypothesis testing stuff!

Moving forward:

- nb day Friday

Next time in lecture:

- Regression! Finally!