CSCI 3022 Intro to Data Science
Small Sample Testing

A summary of our process:

1. State hypothesis: H_0: the baseline or "nothing is interesting result." For the coin: a fair coin, with $p = .5$.
   H_a : what we *want* to test or demonstrate. For the coin: an unfair coin, with $p \neq .5$
2. Collect some data
3. Compute a *test statistic* from our data. Maybe a sample proportion of heads $\hat{p}$?
4. Decide whether the *test statistic* $\hat{p}$ is **too far** from it's assumed baseline value in H_0, and make a decision accordingly. E.g. was $\hat{p}$ *far enough* from $p = .5$ to actually assert that they're different?
5. α is the value that describes the probability of rejecting a null hypotheses *given* that the hypothesis was true.

Announcements and Reminders

HW 5 due Friday

-

PLAN: part HW 6 tomorrow

Post Pract #2 part 1.

Rejection Regions or Probabilities?

How would we know when the test statistic is "sufficiently rare" under the null hypothesis such that we might regard the null as false? We could define a rejection region: a range of values that leads a researcher to reject the null hypothesis.

We can *either:*

1. Define a range of x-values - in the *units* of the data - that correspond to z-values (on the standard normal) that represent "extremely far" from the hypothesized mean. Reject if they're far enough, where far enough is beyond the z_{crit}/t_{crit} value that depends on α
2. Compute a *probability*: if the null hypothesis is true, exactly how "extreme" is our data, as a probability? If it's in the α proportion of most extreme or outlying outcomes when the null is true, maybe we should conclude the null *wasn't* true.

Test for Population Mean (Small Sample)

When the sample size is small and the population is normal, we can use a t-test.
Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

Alternative Hypothesis	Rejection Region for α level test:

Test for Population Mean (Small Sample)

When the sample size is small and the population is normal, we can use a t-test.
Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

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

Alternative Hypothesis Rejection Region for α level test:

Test for Population Mean (Small Sample)

When the sample size is small and the population is normal, we can use a t-test.
Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

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

Alternative Hypothesis	Rejection Region for α level test:
$H_a : \mu > \mu_0$	
$H_a : \mu < \mu_0$	
$H_a : \mu \neq \mu_0$	

Test for Population Mean (Small Sample)

When the sample size is small and the population is normal, we can use a t-test.
Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

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

Alternative Hypothesis	Rejection Region for α level test:
$H_a : \mu > \mu_0$	$t > t_\alpha$
$H_a : \mu < \mu_0$	$t < -t_{\alpha, n-1}$
$H_a : \mu \neq \mu_0$	$t < -t_{\alpha/2, n-1}$ **or** $t > t_{\alpha/2, n-1}$

Comparing 2 Means: Review

The natural estimator of $\mu_1 - \mu_2$ is _____.

CI for $\mu_1 - \mu_2$: $(\bar{X}_1 - \bar{X}_2) \pm z_{C.L.,\alpha_2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Inferential procedures are based on standardizing estimators, so as before we need the mean and standard deviation of _____.

Mean of :

Variance/Standard Deviation of :

Comparing 2 Means: Review

The natural estimator of $\mu_1 - \mu_2$ is $\underline{\bar{X} - \bar{Y}}$.

Inferential procedures are based on standardizing estimators, so as before we need the mean and standard deviation of $\underline{\bar{X} - \bar{Y}}$.

Mean of $\bar{X} - \bar{Y}$:

Variance/Standard Deviation of $\bar{X} - \bar{Y}$:

Comparing 2 Means: Review

The natural estimator of $\mu_1 - \mu_2$ is $\underline{\bar{X} - \bar{Y}}$.

Inferential procedures are based on standardizing estimators, so as before we need the mean and standard deviation of $\underline{\bar{X} - \bar{Y}}$.

Mean of $\bar{X} - \bar{Y}$:

$$E[\bar{X} - \bar{Y}] = E\left[\frac{\sum_i X_i}{n} - \frac{\sum_j Y_j}{m}\right] = \cdots = \mu_1 - \mu_2$$

Variance/Standard Deviation of $\bar{X} - \bar{Y}$:

$$Var[\bar{X} - \bar{Y}] = Var\left[\frac{\sum_i X_i}{n} - \frac{\sum_j Y_j}{m}\right] = Var[\bar{X}] + Var[\bar{Y}] = \ldots$$

$$= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$$

Comparing 2 Means

2 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:

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

Standardizing our estimator gives:

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

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

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

Test Procedures for Normal Populations with Known Variances

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

"μ_1 is Δ_0 greater than μ_2"

Test statistic value:

"$\mu_1 = \mu_2$" if $\Delta_0 = 0$.

Alt Hypothesis	Rejection Region	p-value:

Test Procedures for Normal Populations with Known Variances

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

Test statistic value:

$$z_{stat} = \frac{\left(\bar{X} - \bar{Y}\right) - (\Delta_0)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

Alt Hypothesis Rejection Region p-value:

Test Procedures for Normal Populations with Known Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

$$z_{stat} = \frac{\left(\bar{X} - \bar{Y}\right) - (\Delta_0)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

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 Known Variances

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

Test statistic value:

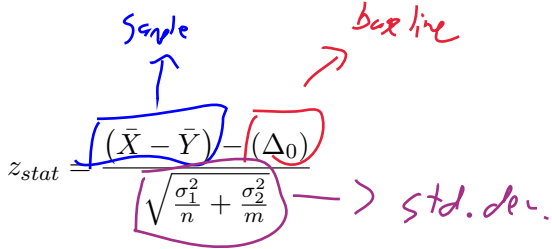
$$z_{stat} = \frac{\left(\bar{X} - \bar{Y}\right) - (\Delta_0)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

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

Test Procedures for Normal Populations with Known Variances

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

Test statistic value:



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

vs α.

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.

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

Comparing 2 Means: Small Sample

Assume

When the population distributions are both normal, the standardized variable

$\in$ Statistics!

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

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

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

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

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

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

Comparing 2 Means: Small Sample

The two-sample t confidence interval for μ_1–μ_2 with confidence level $(1-\alpha)\cdot 100\%$ is then:

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

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)

Test statistic value:

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

Alt Hypothesis	Rejection Region	p-value:

Test Procedures for Normal Populations with Unknown Variances

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

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

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: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

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

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

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:
$H_a : \mu > \mu_0$	$t_{stat} > t_{\alpha,\nu}$	$P(T > t_{stat})$
$H_a : \mu < \mu_0$	$t_{stat} < -t_{\alpha,\nu}$	$P(T < t_{stat})$
$H_a : \mu \neq \mu_0$	$\lvert t_{stat}\rvert > t_{\alpha/2,\nu}$	$P(\lvert T\rvert > \lvert t_{stat}\rvert)$

Test for Equivalence of Proportions

Theoretically, we know that:

2 samples, each is a Y/N or proportion.

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)\left(\frac{1}{n}\right)}}$$

$p = .5$

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:

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

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:

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

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:

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: $\hat{p} = \frac{X+Y}{n+m}$ = combined sample proportion

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: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

Alt Hypothesis	Rejection Region	p-value:

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: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

$$z_{stat} = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

diff data — (0) baseline

std. dev. uses $\hat{p}$ combined

Alt Hypothesis | Rejection Region | p-value:

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: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

$$z_{stat} = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

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: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

$$z_{stat} = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

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

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: $H_0 : \mu_1 - \mu_2 = \Delta_0 \quad (\text{often } 0)$

Test statistic value:

$$z_{stat} = \frac{(\hat{p_1} - \hat{p_2})}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

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

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.

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

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

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \textbf{ OR } \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \textbf{ OR } \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \textbf{ OR } \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

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

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

Now what?

We're almost done talking about CI's and Hypothesis tests. Where are our gaps?

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!
   - 2.3 We will cover 2 more cases of this: variances and bootstrapping.

CLT or properties of $\bar{X}$.

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

"pick itself up by the bootstraps"

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.

repeats ✓ fake data Original data

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

~ each entry an exercise: 40% "2" 20% "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.

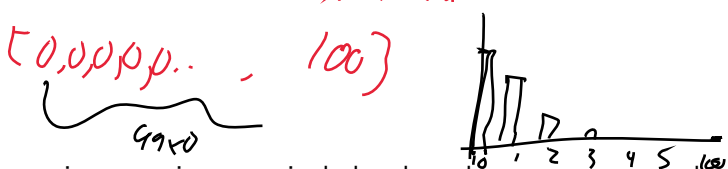
As a rule-of-thumb, each bootstrapped sample should be of the same size as the original sample.

"full"
→ simulate 100 means, 95% CI = MIDDLE 95%

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.

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



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
2. Standard Deviations
3. Other measures that we may not even have theories for!

by practicum

Bootstrapping a median

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

1) 1000 make new "fake" simulation

2) Compute variance of each: 1000 diff s^2

3) Find cutoffs, the .05 & .95 percentiles?
sort the 1000, find the 50th & 950th → [l, u].

Bootstrapping a median

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

1. vars=[]
   nsamp=10000
2. for i in range(nsamp):
   newX=np.random.choice(X, size=len(X), replace=True)
   vars.append(np.var(newX, ddof=1))
3. CI= np.percentile(vars, [5,95])

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

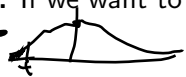
Sometimes we really want to know about various statistics on e.g. the Poisson or Exponential *without* solving some challenging integral or sum or whatever else equations.

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 Pois(λ). Estimate the parameter, e.g. $\lambda \approx \bar{X}$.
2. Simulate a bootstrapped sample from 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!

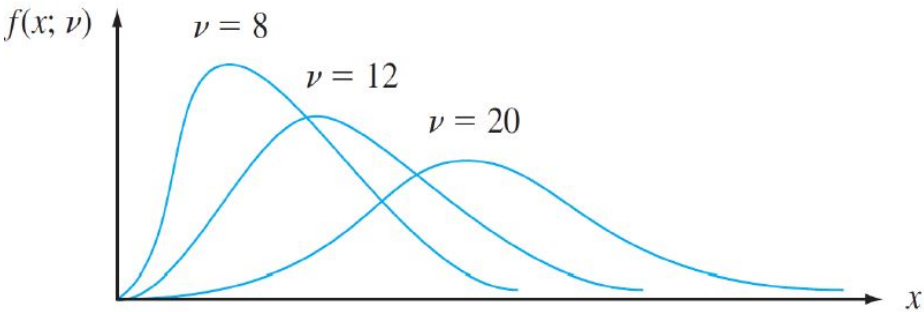
Special Cases: Variance

Definition: *Chi-Squared*
Let ν be a positive integer. The random variable X has a chi-squared distribution with parameter ν if the pdf of X is:

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

The parameter ν is called the number of degrees of freedom (df) of X. The symbol χ^2 is often used in place of "chi-squared."

Special Cases: Variance



Special Cases: Variance

Let $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then the random variable:

has a chi-squared (____) probability distribution with $n - 1$ df.

(In this class, we don't consider the case where the data is not normally distributed.)

Special Cases: Variance

Let $X_1, X_2, \ldots X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Then the random variable:

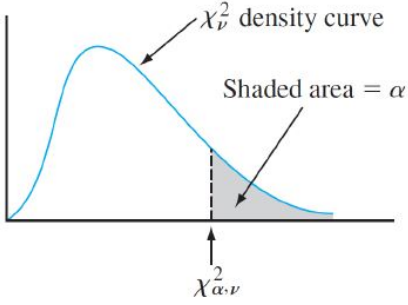
$$\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2}$$

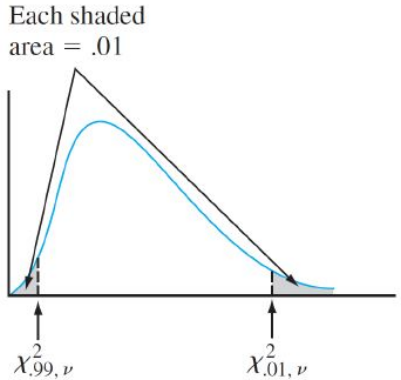
has a chi-squared (χ^2) probability distribution with $n - 1$ df.

(In this class, we don't consider the case where the data is not normally distributed.)

Special Cases: Variance

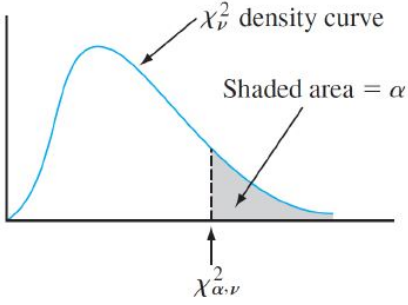
The chi-squared distribution is not symmetric, so these tables and functions contain values of both for near 0 and 1.

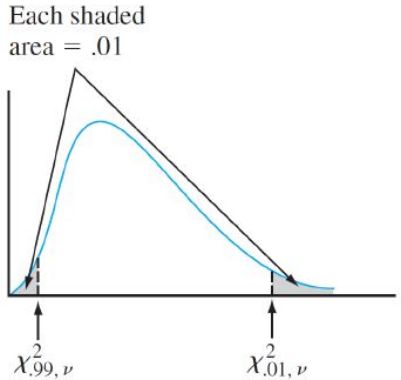




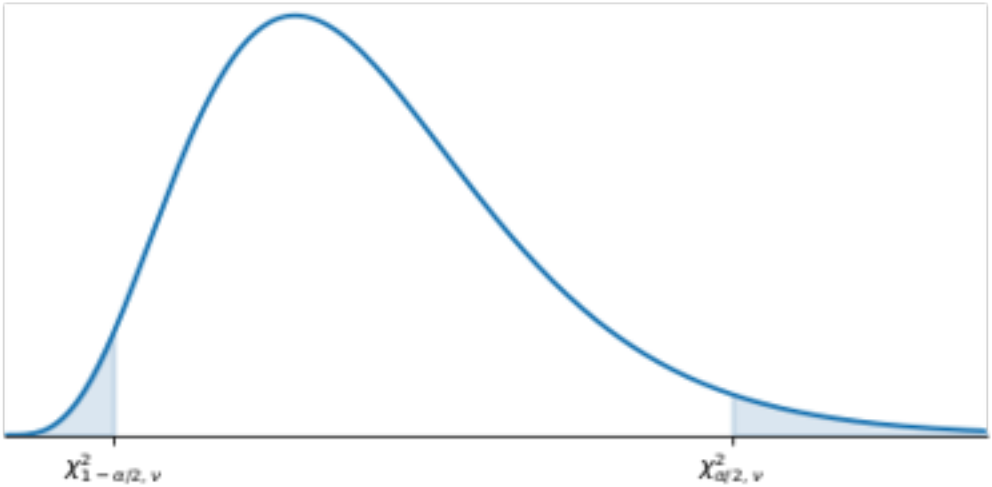
Special Cases: Variance

The chi-squared distribution is not symmetric, so these tables and functions contain values of χ^2_α both for near 0 and 1.





Two tailed χ^2



Special Cases: Variance

As a consequence:

$$1 - \alpha = P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right)$$

Or, equivalently:

Thus we have a confidence interval for the variance. Taking square roots gives a CI for the standard deviation.

Special Cases: Variance

As a consequence:

$$1-\alpha = P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right)$$

$$= P(1/\chi^2_{1-\alpha/2,n-1} \geq \frac{\sigma^2}{(n-1)s^2} \geq 1/\chi^2_{\alpha/2,n-1})$$

Or, equivalently:

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

is a $100\%(1-\alpha)$ CI for σ^2.
Thus we have a confidence interval for the variance. Taking square roots gives a CI for the standard deviation.

A CI on Variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation is too large. She selected 10 bags at random and weights them, for a sample variance of $4.2g^2$. Find a 95% CI for the variance and a 95% CI for the SD.

$\alpha = .05, \qquad \alpha/2 = .025 \qquad n = 10 \qquad s^2 = 4.2$

A CI on Variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation is too large. She selected 10 bags at random and weights them, for a sample variance of $4.2g^2$. Find a 95% CI for the variance and a 95% CI for the SD.

$\alpha = .05, \qquad \alpha/2 = .025 \qquad n = 10 \qquad s^2 = 4.2$

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{.975,9} = \texttt{stats.chi2.ppf(0.025,9)} = 2.70$$

$$\chi^2_{\alpha/2,n-1} = \chi^2_{.025,9} = \texttt{stats.chi2.ppf(0.975,9)} = 19.02$$

A CI on Variance

Example: A large candy manufacturer produces packages of candy targeted to weigh 52g. The weight of the packages of candy is known to be normally distributed, but a QC engineer is concerned that the variation is too large. She selected 10 bags at random and weights them, for a sample variance of $4.2g^2$. Find a 95% CI for the variance and a 95% CI for the SD.

$\alpha = .05, \qquad \alpha/2 = .025 \qquad n = 10 \qquad s^2 = 4.2$

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{.975,9} = \texttt{stats.chi2.ppf(0.025,9)} = 2.70$$

$$\chi^2_{\alpha/2,n-1} = \chi^2_{.025,9} = \texttt{stats.chi2.ppf(0.975,9)} = 19.02$$

$$\frac{(10-1)4.2}{19.02} < \sigma^2 \frac{(10-1)4.2}{2.70} \implies 1.99 < \sigma^2 < 14.0$$

$$\implies \sqrt{1.99} < \sigma < \sqrt{14.0}$$

Test for Equivalence of Variance

The F probability distribution has two parameters, denoted by ν_1 and ν_2. The parameter ν_1 is called the numerator degrees of freedom, and ν_2 is the denominator degrees of freedom.

A random variable that has an F distribution cannot assume a negative value. The density function is complicated and will not be used explicitly, so it's not shown.

There is an important connection between an F variable and chisquared variables.

Test for Equivalence of Variance

If X_1 and X_2 are independent chi-squared rv's with ν_1 and ν_2 df, respectively, then the rv

can be shown to have an F distribution.

Recall that a chi-squared distribution was obtain by summing squared standard Normal variables (such as squared deviations for example). So a scaled ratio of two variances is a ratio of two scaled chi-squared variables.

Test for Equivalence of Variance

If X_1 and X_2 are independent chi-squared rv's with ν_1 and ν_2 df, respectively, then the rv

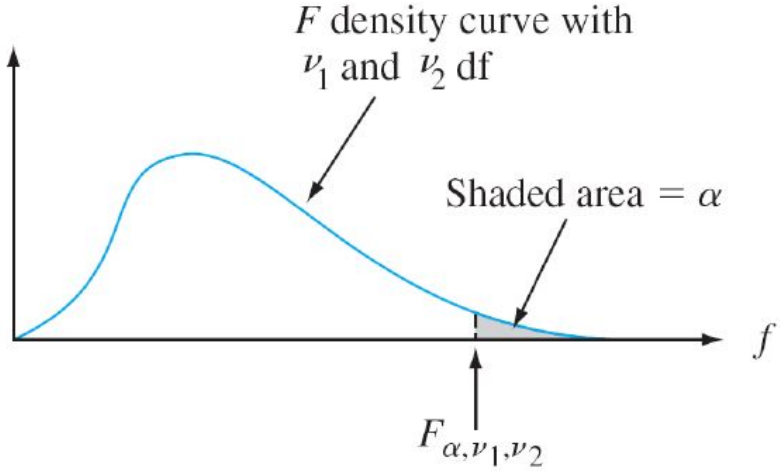
$$F = \frac{X_1/\nu_1}{X_2/\nu_2}$$

can be shown to have an F distribution.

Recall that a chi-squared distribution was obtain by summing squared standard Normal variables (such as squared deviations for example). So a scaled ratio of two variances is a ratio of two scaled chi-squared variables.

Test for Equivalence of Variance

Figure below illustrates a typical F density function.:



Test for Equivalence of Variance

We use F_{α,ν_1,ν_2} for the value on the horizontal axis that captures of the area under the F density curve with ν_1 and ν_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,\nu_1,\nu_2} =$$

Test for Equivalence of Variance

We use F_{α,ν_1,ν_2} for the value on the horizontal axis that captures of the area under the F density curve with ν_1 and ν_2 df in the upper tail.

The density curve is not symmetric, so it would seem that both upper- and lower-tail critical values must be tabulated. This is not necessary, though, because of the fact that

$$F_{1-\alpha,\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_1,\nu_2}}$$

For example, $F_{.05,6,10} = 3.22$ and $F_{.95,10,6} = 0.31 = 1/3.22$.

Test for Equivalence of Variance

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem:

Let $X_1, X_2, \ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2 let $Y_1, Y_2, \ldots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances. Then the rv

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

Test for Equivalence of Variance

A test procedure for hypotheses concerning the ratio σ_1^2/σ_2^2 is based on the following result.

Theorem:

Let $X_1, X_2, \ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2 let $Y_1, Y_2, \ldots, Y_n$ be another random sample (independent of the X_i's) from a normal distribution with variance σ_2^2 and let s_1^2 and s_2^2 denote the two sample variances. Then the rv

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has an F distribution with $\nu_1 = m - 1$ and $\nu_2 = n - 1$.

Test for Equivalence of Variance

This theorem results from combining the fact that the variables $\frac{(n-1)s_2^2}{\sigma_2^2}$ and $\frac{(m-1)s_1^2}{\sigma_1^2}$ each have a chi-squared distribution with $n-1$ and $m-1$ df, respectively.

Because F involves a ratio rather than a difference, the test statistic is the ratio of sample variances.

The claim that $\sigma_1^2 = \sigma_2^2$ is then rejected if the ratio s_1^2/s_2^2 differs by too much from 1.

Test for Equivalence of Variance

Null hypothesis: H_0 :

Test statistic value:

Alt Hypothesis	Rejection Region	p-value:

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis	Rejection Region	p-value:

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \sigma_1^2 > \sigma_2^2$		
$H_a : \sigma_1^2 < \sigma_2^2$		
$H_a : \sigma_1^2 \neq \sigma_2^2$		

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \sigma_1^2 > \sigma_2^2$	$F_{stat} > F_{\alpha,m-1,n-1}$	
$H_a : \sigma_1^2 < \sigma_2^2$	$F_{stat} < F_{1-\alpha,m-1,n-1}$	
$H_a : \sigma_1^2 \neq \sigma_2^2$	$F_{stat} < F_{1-\alpha/2,m-1,n-1}$ OR $F_{stat} > F_{\alpha/2,m-1,n-1}$	

Test for Equivalence of Variance

Null hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$

Test statistic value:

$$F = s_1^2 / s_2^2$$

Alt Hypothesis	Rejection Region	p-value:
$H_a : \sigma_1^2 > \sigma_2^2$	$F_{stat} > F_{\alpha,m-1,n-1}$	$P(F_{m-1,n-1} > F_{stat})$
$H_a : \sigma_1^2 < \sigma_2^2$	$F_{stat} < F_{1-\alpha,m-1,n-1}$	$P(F_{m-1,n-1} < F_{stat})$
$H_a : \sigma_1^2 \neq \sigma_2^2$	$F_{stat} < F_{1-\alpha/2,m-1,n-1}$ OR $F_{stat} > F_{\alpha/2,m-1,n-1}$	(OR)

Test for Equivalence of Variance

Example: On the basis of data reported in the article "Serum Ferritin in an Elderly Population" (J. of Gerontology, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men? Use alpha $= .01$.

Test for Equivalence of Variance

Example: On the basis of data reported in the article "Serum Ferritin in an Elderly Population" (J. of Gerontology, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men? Use alpha = .01.

$$F_{27,25} = \frac{52.6^2}{84.2^2} = F_{stat}$$

Test for Equivalence of Variance

Example: On the basis of data reported in the article "Serum Ferritin in an Elderly Population" (J. of Gerontology, 1979: 521–524), the authors concluded that the ferritin distribution in the elderly had a smaller variance than in the younger adults. (Serum ferritin is used in diagnosing iron deficiency.)

For a sample of 28 elderly men, the sample standard deviation of serum ferritin (mg/L) was $s_1 = 52.6$; for 26 young men, the sample standard deviation was $s_2 = 84.2$.

Does this data support the conclusion as applied to men? Use alpha = .01.

$$F_{27,25} = \frac{52.6^2}{84.2^2} = F_{stat}$$

$$P(F_{27,25} \leq \frac{52.6^2}{84.2^2}) = \texttt{stats.f.cdf}(\frac{52.6^2}{84.2^2}, 27, 25) = 0.0093 = p < \alpha = 0.01$$

Now what?

On to Regression!

A few things to note: you are **not expected** nor even encouraged to memorize all of these formulas. Instead, you want a few basic vocabulary words and associations:

1. **Normals** are for large sample measures of the *mean* (or proportions). They are (difference)/(standard deviation) formulas.
2. **t's** are for small sample measures of the *mean*. *Assumption:* populations are normal. They are (difference)/(standard deviation) formulas.
3. **Chi-squared** are for measures of the variance. *Assumption:* of a normal. They use (sums of squared deviations) in the formula.
4. **F** are for measures of the variance. *Assumption:* of a normal. They are a *ratio* of two variances/chi-squareds.

With those associations and basic algebraic intuitions, just look up the one you need at any given time!

Daily Recap

Today we learned

1. Intro and Basics of Hypothesis Tests

Moving forward:

- nb day Friday for HTs

Next time in lecture:

- More Hypotheses!