Quantitative Methods in Finance

Step 1: Explore

Prof. Mike Aguilar

https:/www.linkedin.com/in/mike-aguilar-econ

Last updated: March 19, 2024

Audience and Prerequisites

Audience:

- These notes were prepared for Kenan Flagler Business School's Daytime MBA program
- The setting is a 14 session course
- These notes serve as reference materials to complement our in-class work using computational tools

Prerequisites:

- Some knowledge of probability, statistics are expected
- Knowledge of finance in general, and asset pricing, in particular is expected

Motivating Case Study

- The motivating case study is portfolio allocation
- However, the concepts and tools are widely applicable to a range of settings within Finance
- We group the concepts into the following functional steps within our case study
  1. Explore - Loading and cleaning data, EDA, etc..
  2. Explain - Factor modeling, etc..
  3. Forecast - Time series models, etc...
  4. Protect - Portfolio allocation, risk measurement, etc..

Housekeeping

- Throughout the slide deck you will see “Q”, which indicates a question to you, the reader
- You will also see “A”, which indicates the associated answer
- It is generally most efficient to learn this material through active participation. Whenever you encounter a “Q”, be sure to try and develop your answer before turning the page to the provided “A” answer
- I recommend reading through these slides before engaging in associated coding exercises

Outline

Defining Returns

- Realized return $\approx$ Cash Flows over life of investment as a % of purchase price.
- Unrealized returns are changes in asset value, which have yet to be converted into cash.
- For a stock, the cash flows are dividends. Unrealized returns are capital gains, i.e. a change in the stock's price.
- For a bond, the cash flows might be interest and face value. Unrealized returns are similarly a change in the bond's market price.

Outline

Definition (Holding Period Yield)

Consider an asset without intermediate cash flows. Set t_0 as the time of purchase, and t_1 as the time of sale.

$$R(t_0, t_1) = \frac{P_{t_1} - P_{t_0}}{P_{t_0}} = \frac{P_{t_1}}{P_{t_0}} - 1$$

Definition (Holding Period Yield)

Consider an asset without intermediate cash flows. Set t_0 as the time of purchase, and t_1 as the time of sale.

$$R(t_0, t_1) = \frac{P_{t_1} - P_{t_0}}{P_{t_0}} = \frac{P_{t_1}}{P_{t_0}} - 1$$

Definition (Holding Period Return)

Consider an asset without intermediate cash flows. Set t_0 as the time of purchase, and t_1 as the time of sale.

$$1 + R(t_0, t_1) = \frac{P_{t_1}}{P_{t_0}}$$

Note: the term "return" is used colloquially for both HPY and HPR.

Definition (Simple Net Return)

Assume 1 day holding period. Then we can define a one day simple net return as

$$R_t = \frac{P_t}{P_{t-1}} - 1$$

Definition (Simple Net Return)

Assume 1 day holding period. Then we can define a one day simple net return as

$$R_t = \frac{P_t}{P_{t-1}} - 1$$

Definition (Simple Gross Return)

Assume 1 day holding period. Then we can define a one day gross return as

$$1 + R_t = \frac{P_t}{P_{t-1}}$$

Multi-Period Returns

The 2-day return (HPY) can be computed as

$$R_{t,t-2} \equiv R_t(2) = \frac{P_t - P_{t-2}}{P_{t-2}} = \frac{P_t}{P_{t-2}} - 1$$

We can write this as

Multi-Period Returns

The 2-day return (HPY) can be computed as

$$R_{t,t-2} \equiv R_t(2) = \frac{P_t - P_{t-2}}{P_{t-2}} = \frac{P_t}{P_{t-2}} - 1$$

We can write this as

$$\begin{aligned} R_t(2) &= \frac{P_t}{P_{t-1}} \frac{P_{t-1}}{P_{t-2}} - 1 \\ &= (1 + R_t)(1 + R_{t-1}) - 1 \\ &= R_{t-1} + R_t + R_t R_{t-1} \\ R_t(2) &\approx R_{t-1} + R_t \end{aligned}$$

if R_t and R_{t-1} are small so that $R_t R_{t-1} \approx 0$.

Multi-Period Returns

Definition

Consider an asset without intermediate cash flows. The H-period simple net return is

$$R_t(H) \approx \sum_{j=0}^{H-1} R_{t-j}$$

Multi-Period Returns

Definition

Consider an asset without intermediate cash flows. The H-period simple net return is

$$R_t(H) \approx \sum_{j=0}^{H-1} R_{t-j}$$

Definition

Consider an asset without intermediate cash flows. The H-period simple gross return (aka Geometric Return) is

$$\begin{aligned} 1 + R_t(H) &= (1 + R_t)(1 + R_{t-1}) \cdots (1 + R_{t-H+1}) \\ &= \prod_{j=0}^{H-1} (1 + R_{t-j}) \end{aligned}$$

Stock Returns

The intermediate cash flows for a stock are the dividends. We can adjust return calculations as follows:

$$
\begin{aligned}
R_t &= \frac{P_t + D_t - P_{t-1}}{P_{t-1}} \\
&= \frac{P_t + D_t}{P_{t-1}} - 1 \\
1 + R_t &= \frac{P_t}{P_{t-1}} + \frac{D_t}{P_{t-1}}
\end{aligned}
$$

Stock Returns

The intermediate cash flows for a stock are the dividends. We can adjust return calculations as follows:

$$
\begin{aligned}
R_t &= \frac{P_t + D_t - P_{t-1}}{P_{t-1}} \\
&= \frac{P_t + D_t}{P_{t-1}} - 1 \\
1 + R_t &= \frac{P_t}{P_{t-1}} + \frac{D_t}{P_{t-1}}
\end{aligned}
$$

$$\text{Stock Returns} = \text{Capital Gains Yield} + \text{Dividend Yield}$$

Bond Returns

The 1yr holding period for a 30yr US Treasury bond with a 6% annual coupon rate is

$$
\begin{aligned}
R_t &= \frac{P_t + C_t - P_{t-1}}{P_{t-1}} \\
&= \frac{P_t + C_t}{P_{t-1}} - 1 \\
1 + R_t &= \frac{P_t}{P_{t-1}} + \frac{C_t}{P_{t-1}}
\end{aligned}
$$

Bond Returns

The 1yr holding period for a 30yr US Treasury bond with a 6% annual coupon rate is

$$\begin{aligned} R_t &= \frac{P_t + C_t - P_{t-1}}{P_{t-1}} \\ &= \frac{P_t + C_t}{P_{t-1}} - 1 \\ 1 + R_t &= \frac{P_t}{P_{t-1}} + \frac{C_t}{P_{t-1}} \end{aligned}$$

$$\text{Bond Returns} = \text{Capital Gains Yield} + \text{Coupon Rate}$$

Assumes purchased at par such that P_{t-1} equals face value.

Outline

1. Compute Returns
   - Simple Returns
     - Annualized Returns

Annualized Returns

Example

Consider an asset measured at a quarterly frequency. Suppose the most recent quarterly return is $R_t = \frac{P_t}{P_{t-1}} - 1$. What is the annualized rate of return on this asset (R_A)?

Annualized Returns

Example

Consider an asset measured at a quarterly frequency. Suppose the most recent quarterly return is $R_t = \frac{P_t}{P_{t-1}} - 1$. What is the annualized rate of return on this asset (R_A)?

We assume that the asset will continue to grow at its current pace for the rest of the year, such that $R_t = R_{t+1} = R_{t+2} = R_{t+3} = R$. Then, $1 + R_t(4) = (1 + R_t)(1 + R_{t+1})(1 + R_{t+2})(1 + R_{t+3}) = (1 + R)^4$. This implies that $R^A(1) = (1 + R)^4 - 1$.

Definition (Simple Annualized Returns)

$$R_t(H)^A = (1 + R_t(H))^{\frac{\#\text{of periods per yr}}{H}} - 1$$

$$R_t(H)^A = \left(\frac{P_t}{P_{t-H}}\right)^{\frac{\#\text{of periods per yr}}{H}} - 1$$

Outline

1. Compute Returns
   - Simple Returns
   - Inflation Adjusted Returns
   - Continuously Compounded Returns
   - Log vs Simple

Real Returns

To adjust for price changes, macroeconomists give us the relation:

$$\text{Real} = \frac{\text{Nominal}}{\text{Deflator}}$$

Let's use the CPI as our measure of the price level, and inflation given by

$$\pi_t = \frac{CPI_t}{CPI_{t-1}} - 1$$

We can define P^{real} as the (real) inflation adjusted price, such that

$$P_t^{real} = \frac{P_t}{CPI_t}$$

Real Returns

We can "deflate" returns by

$$R_t^{real} = \frac{P_t^{real}}{P_{t-1}^{real}} - 1 = \frac{\frac{P_t}{CPI_t}}{\frac{P_{t-1}}{CPI_{t-1}}} - 1$$

$$= \frac{P_t}{P_{t-1}} \frac{CPI_{t-1}}{CPI_t} - 1$$

Notice: $\frac{CPI_t}{CPI_{t-1}} = 1 + \pi_t$.

Real Returns

We can "deflate" returns by

$$R_t^{real} = \frac{P_t^{real}}{P_{t-1}^{real}} - 1 = \frac{\frac{P_t}{CPI_t}}{\frac{P_{t-1}}{CPI_{t-1}}} - 1$$

$$= \frac{P_t}{P_{t-1}} \frac{CPI_{t-1}}{CPI_t} - 1$$

Notice: $\frac{CPI_t}{CPI_{t-1}} = 1 + \pi_t$.

Definition (Simple Inflation Adjusted Returns)

$$R_t^{real} = \frac{1 + R_t}{1 + \pi_t} - 1$$

Outline

Calculating Domestic Currency Returns

Calculating Domestic Currency Returns

1. Start with (for instance) \$100.
2. Convert to euros with $FX_{t-1} = \frac{e_{t-1}\$}{€} \rightarrow \frac{\$100}{FX_{t-1}}$
3. Grow at euro rate: $\frac{\$100}{FX_{t-1}}(1 + R_t^{€})$
4. Repatriate at $FX_t \rightarrow \frac{\$100}{FX_{t-1}}(1 + R_t^{€})FX_t$
5. Compute dollar returns $\frac{\frac{\$100}{FX_{t-1}}(1+R_t^{€})FX_t}{\$100} - 1$

Calculating Domestic Currency Returns

1. Start with (for instance) \$100.
2. Convert to euros with $FX_{t-1} = \frac{e_{t-1}\$}{€} \rightarrow \frac{\$100}{FX_{t-1}}$
3. Grow at euro rate: $\frac{\$100}{FX_{t-1}}(1 + R_t^{€})$
4. Repatriate at $FX_t \rightarrow \frac{\$100}{FX_{t-1}}(1 + R_t^{€})FX_t$
5. Compute dollar returns $\frac{\frac{\$100}{FX_{t-1}}(1+R_t^{€})FX_t}{\$100} - 1$

Definition (Currency Adjusted Returns)

Denote \$ as the domestic currency and € as the foreign currency. Define the exchange rate as $FX_t = \frac{e_t\$}{€}$. Then the one period dollar-return of an investment made in a euro denominated asset can be written as

$$R_t^{\$} = \frac{FX_t}{FX_{t-1}}(1 + R_t^{€}) - 1$$

Outline

1. Compute Returns
   - Simple Returns
   - Inflation Adjusted Returns
   - Continuously Compounded Returns
   - Log vs Simple

Definition (Log Returns)

Consider an asset with no intermediate cash flows. Define the *ln* operator as the natural log; i.e. base e, not base 10. The one period continuously compounded return is

$$r_t = ln(1 + R_t) = ln(\frac{P_t}{P_{t-1}}) = ln(P_t) - ln(P_{t-1})$$

Note that we can always convert from continuously compounded to simple returns:

$$R_t = e^{r_t} - 1$$

Multi-period Returns

$$
\begin{aligned}
r_t(2) &= ln(1 + R_t(2)) = ln\left(\frac{P_t}{P_{t-2}}\right) \\
&= ln\left(\frac{P_t}{P_{t-1}}\frac{P_{t-1}}{P_{t-2}}\right) \\
&= ln\left(\frac{P_t}{P_{t-1}}\right) + ln\left(\frac{P_{t-1}}{P_{t-2}}\right) \\
&= r_t + r_{t-1}
\end{aligned}
$$

Multi-period Returns

$$\begin{aligned} r_t(2) &= ln(1 + R_t(2)) = ln\left(\frac{P_t}{P_{t-2}}\right) \\ &= ln\left(\frac{P_t}{P_{t-1}}\frac{P_{t-1}}{P_{t-2}}\right) \\ &= ln\left(\frac{P_t}{P_{t-1}}\right) + ln\left(\frac{P_{t-1}}{P_{t-2}}\right) \\ &= r_t + r_{t-1} \end{aligned}$$

Definition

Consider an asset with no intermediate cash flows. The H-period continuously compounded return is

$$r_t(H) = \sum_{j=0}^{H-1} r_{t-j}$$

Outline

Definition (Continuously Compounded Annualized Returns)

$$r_t(H)^A = \left(\frac{\#\text{of periods per year}}{H}\right) \times r_t(H)$$

Outline

Real Returns

Deflating continuously compounded returns follows a similar logic to simple returns.

$$
\begin{aligned}
r_t^{real} &= ln(1 + R_t^{real}) \\
&= ln\left(\frac{P_t}{P_{t-1}}\frac{CPI_{t-1}}{CPI_t}\right) \\
&= ln\left(\frac{P_t}{P_{t-1}}\right) + ln\left(\frac{CPI_{t-1}}{CPI_t}\right) \\
&= ln\left(\frac{P_t}{P_{t-1}}\right) - ln\left(\frac{CPI_t}{CPI_{t-1}}\right) \\
&= r_t - \pi_t^c
\end{aligned}
$$

where $\pi_t^c = ln(1 + \pi_t)$.

Real Returns

Deflating continuously compounded returns follows a similar logic to simple returns.

$$
\begin{aligned}
r_t^{real} &= ln(1 + R_t^{real}) \\
&= ln\left(\frac{P_t}{P_{t-1}} \frac{CPI_{t-1}}{CPI_t}\right) \\
&= ln\left(\frac{P_t}{P_{t-1}}\right) + ln\left(\frac{CPI_{t-1}}{CPI_t}\right) \\
&= ln\left(\frac{P_t}{P_{t-1}}\right) - ln\left(\frac{CPI_t}{CPI_{t-1}}\right) \\
&= r_t - \pi_t^c
\end{aligned}
$$

where $\pi_t^c = ln(1 + \pi_t)$.

Definition (Continuously Compounded Inflation Adjusted Returns)

$$r_t^{real} = r_t - \pi_t^c$$

Outline

Calculating Domestic Currency Returns

Calculating Domestic Currency Returns

1. Start with (for instance) \$100.
2. Convert to euros with $FX_{t-1} = \frac{e_{t-1}\$}{€} \rightarrow \frac{\$100}{FX_{t-1}}$
3. Grow at euro rate: $\frac{\$100}{FX_{t-1}}(1 + r_t^{€})$
4. Repatriate at $FX_t \rightarrow \frac{\$100}{FX_{t-1}}(1 + r_t^{€})FX_t$
5. Compute dollar returns $\frac{\frac{\$100}{FX_{t-1}}(1+r_t^{€})FX_t}{\$100} - 1$

Calculating Domestic Currency Returns

1. Start with (for instance) \$100.
2. Convert to euros with $FX_{t-1} = \frac{e_{t-1}\$}{€} \rightarrow \frac{\$100}{FX_{t-1}}$
3. Grow at euro rate: $\frac{\$100}{FX_{t-1}}(1 + r_t^{€})$
4. Repatriate at $FX_t \rightarrow \frac{\$100}{FX_{t-1}}(1 + r_t^{€})FX_t$
5. Compute dollar returns $\frac{\frac{\$100}{FX_{t-1}}(1+r_t^{€})FX_t}{\$100} - 1$

Definition (Currency Adjusted Returns)

Denote \$ as the domestic currency and € as the foreign currency. Define the exchange rate as $FX_t = \frac{e_t\$}{€}$. Then the one period dollar-return of an investment made in a euro denominated asset can be written as

$$r_t^{\$} = \frac{FX_t}{FX_{t-1}}(1 + r_t^{€}) - 1$$

Outline

1. Compute Returns
   - Simple Returns
   - Inflation Adjusted Returns
   - Continuously Compounded Returns
   - Log vs Simple

Log Returns or Simple Returns?

Simple Returns

Sum(Logs) $\neq$ Log(Sums), so

Simple Returns are more appropriate for portfolio construction and performance reporting

Log Returns

Time Series properties of sums are easier to analyze than products, so

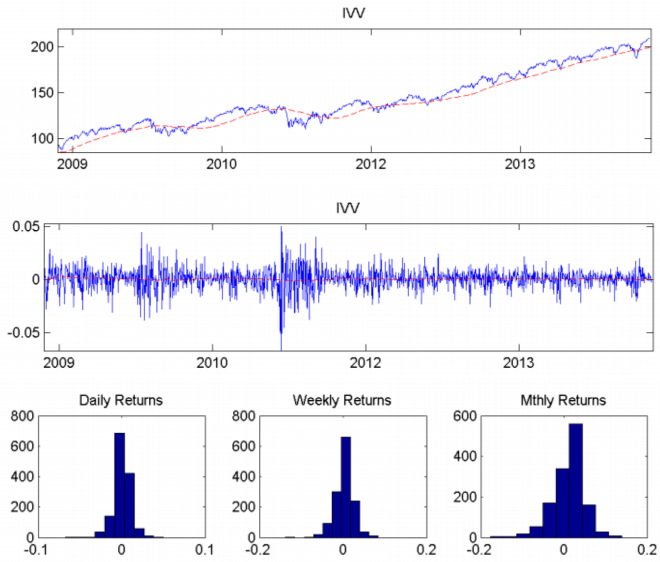
Log Returns are more appropriate for regression and inference

Bottom Line

Simple and Log Returns should be close in most cases.

You need to know when to use each.

Outline


IVV
200
150
100
2009
2010
2012
2013
IVV
0.05
0
-0.05
2009
2010
2012
2013
Daily Returns
800
600
400
200
0
-0.1
0
0.1
Weekly Returns
800
600
400
200
0
-0.2
0
0.2
Mthly Returns
600
400
200
0
-0.2
0
0.2

Outline

2 Describe Returns

- Probability
- Central Tendency, Dispersion, Relatedness
- Common Probability Distributions

What is the probability of an event occurring?

Definition (Probability Density Function (p.d.f.) of X)

Define X as a random variable and x as the realization of that random variable. The pdf of X is defined as $f(x) \geq 0 \; \forall \; x$ such that

$$Pr(a < X \leq b) = \int_a^b f(x)dx$$

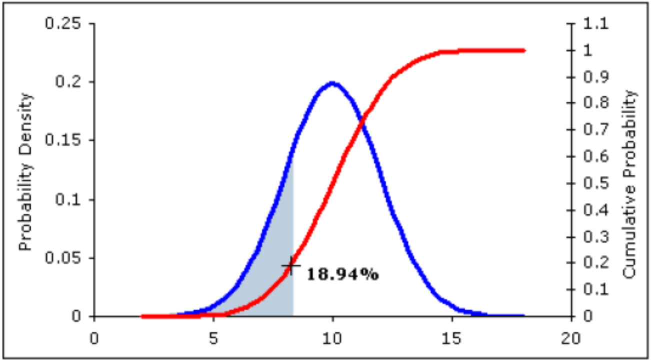
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Definition (Cumulative Distribution Function (c.d.f.) of X)

Define the cdf of X as $F(x)$ such that

$$F(x) = Pr(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

$$F'(x) = \frac{\partial F(x)}{\partial x} = f(x)$$



What is the probability of an event occurring?

Definition (Quantile Function "Inverse CDF")

What is the probability of an event occurring?

Definition (Quantile Function "Inverse CDF")

The value $F^{-1}(p)$ is called the p-quantile of X for each $0 < p < 1$.

What is the probability of an event occurring?

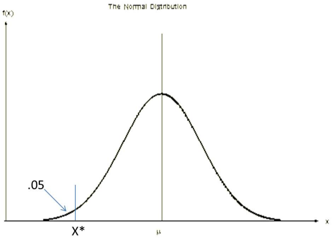
Definition (Quantile Function "Inverse CDF")

The value $F^{-1}(p)$ is called the p-quantile of X for each $0 < p < 1$.

5% of the values of X are less than X^*.

$F(X^*) = Pr(X \leq X^*) = 5\%$.

X^* is also referred to as the 5^{th} percentile



What is the probability of an event occurring?

Commonly Used Quantiles

- Percentiles: .01-quantile, .02-quantile, ..., .99-quantile
- Quintiles: .2-quantile, .4-quantile, .6-quantile, .8-quantile
- Quartiles: .25-quantile, .50-quantile, .75-quantile
  - Interquartile Range: 3rd Quartile - 1st Quartile
- Deciles: .1-quantile, .2-quantile, ..., .9-quantile

Outline

2. Describe Returns
   - Probability
   - Central Tendency, Dispersion, Relatedness
   - Common Probability Distributions

How do we Describe a Probability Distribution?

How do we Describe a Probability Distribution?

Moments of the Distribution

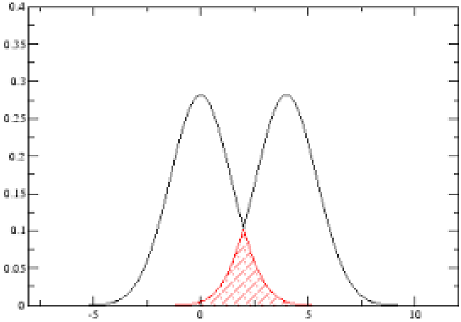
- First Moment = Expected Value [Central Tendency]
- Second Moment = Variance [Dispersion]
- Third Moment = Skewness [Symmetry]
- Fourth Moment = Kurtosis [Tail Thickness]

What is the Central Tendency of the returns?

Definition (Expected Value; i.e. "Population Mean")

Weighted average of all possible values, taking the probability of each outcome as the weights.

$$E(X) \equiv \mu_X = \int_{-\infty}^{+\infty} X f(X) dX$$

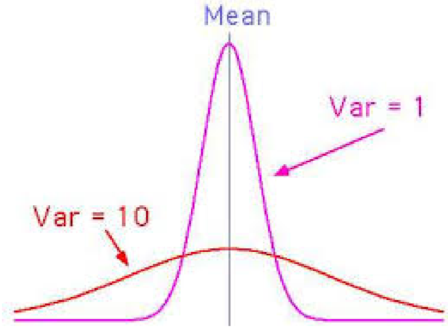


What is the Dispersion of the returns?

Definition (Population Variance)

$$Var(X) \equiv \sigma_X^2 = E\left\{(X-\mu_X)^2\right\} = \int_{-\infty}^{+\infty} (X-\mu_X)^2 f(X) dX$$

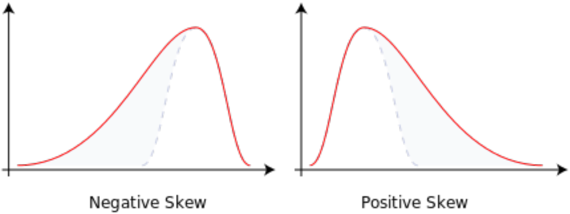
$$Std(X) \equiv \sigma_X = \sqrt{\sigma_X^2}$$



Are returns symmetric?

Definition (Population Skewness)

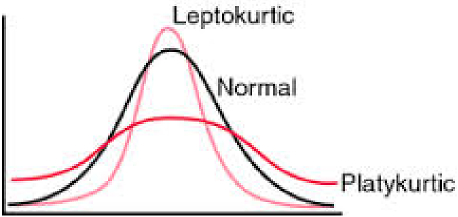
$$S(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



Are returns heavy-tailed?

Definition (Population Kurtosis)

$$K(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$



What is the Relationship between Two R.V.s?

Definition (Covariance & Correlation)

$$Cov(X,Y) \equiv \sigma_{XY} = E\{(X-\mu_X)(Y-\mu_Y)\} = E[XY] - \mu_X\mu_Y$$

$$Corr(X,Y) \equiv \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

What is the Relationship between Two R.V.s?

Definition (Covariance & Correlation)

$$
\begin{aligned}
Cov(X,Y) \equiv \sigma_{XY} &= E\{(X-\mu_X)(Y-\mu_Y)\} = E[XY] - \mu_X\mu_Y \\
Corr(X,Y) \equiv \rho_{XY} &= \frac{\sigma_{XY}}{\sigma_X\sigma_Y}
\end{aligned}
$$

Definition (Independence)

Random variables X and Y are independent if
$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for any $g(X)$ and $h(Y)$.

- Independence implies covariance $= 0$.
- But covariance $= 0$ does not imply independence.

What is the Relationship between Two R.V.s?

Definition (Independent and Identically Distributed (iid))

Two random variables are i.i.d if they have the same distribution (i.e. same family as well as moments) and are independent.

Common Estimators of Moments

Population	Estimator (i.e. "Sample")

Common Estimators of Moments

Population	Estimator (i.e. "Sample")
$E[X] \equiv \mu$	$\bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t$

Common Estimators of Moments

Population	Estimator (i.e. "Sample")
$E[X] \equiv \mu$	$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$
σ_X^2	$s_X^2 = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})^2; s_X = \sqrt{s_X^2}$

Common Estimators of Moments

Population	Estimator (i.e. "Sample")
$E[X] \equiv \mu$	$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$
σ_X^2	$s_X^2 = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})^2; s_X = \sqrt{s_X^2}$
$S(X)$	$\hat{S}(X) = \frac{T}{(T-1)(T-2)}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$ $\hat{S}(X) \approx \frac{1}{T}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$

Common Estimators of Moments

Population	Estimator (i.e. "Sample")
$E[X] \equiv \mu$	$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$
σ_X^2	$s_X^2 = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})^2; s_X = \sqrt{s_X^2}$
$S(X)$	$\hat{S}(X) = \frac{T}{(T-1)(T-2)} \frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$ $\hat{S}(X) \approx \frac{1}{T} \frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$
$K(X)$	$\hat{K}(X) = \frac{T(T+1)}{(T-1)(T-2)(T-3)} \frac{\sum_{t=1}^{T}(X_t-\bar{X})^4}{s_X^4}$ $\hat{K}(X) \approx \frac{1}{t} \frac{\sum_{t=1}^{T}(X_t-\bar{X})^4}{s_X^4}$

Common Estimators of Moments

Population	Estimator (i.e. "Sample")
$E[X] \equiv \mu$	$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$
σ_X^2	$s_X^2 = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})^2; s_X = \sqrt{s_X^2}$
$S(X)$	$\hat{S}(X) = \frac{T}{(T-1)(T-2)}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$ $\hat{S}(X) \approx \frac{1}{T}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$
$K(X)$	$\hat{K}(X) = \frac{T(T+1)}{(T-1)(T-2)(T-3)}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^4}{s_X^4}$ $\hat{K}(X) \approx \frac{1}{t}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^4}{s_X^4}$
σ_{XY}	$s_{XY} = \frac{1}{T-1}\sum_{T=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})$

Common Estimators of Moments

Population	Estimator (i.e. "Sample")
$E[X] \equiv \mu$	$\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$
σ_X^2	$s_X^2 = \frac{1}{T-1}\sum_{t=1}^{T}(X_t - \bar{X})^2; s_X = \sqrt{s_X^2}$
$S(X)$	$\hat{S}(X) = \frac{T}{(T-1)(T-2)}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$ $\hat{S}(X) \approx \frac{1}{T}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^3}{s_X^3}$
$K(X)$	$\hat{K}(X) = \frac{T(T+1)}{(T-1)(T-2)(T-3)}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^4}{s_X^4}$ $\hat{K}(X) \approx \frac{1}{t}\frac{\sum_{t=1}^{T}(X_t-\bar{X})^4}{s_X^4}$
σ_{XY} ρ_{XY}	$s_{XY} = \frac{1}{T-1}\sum_{T=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})$ $r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}}$

Arithmetic vs Geometric Mean Return

When applied to returns, the sample estimator $\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t$ is referred to as the "arithmetic" mean return.

	Arithmetic	Geometric
Formula	$\frac{1}{T}\sum_{t=1}^{T} R_t$	$[\prod_{t=1}^{T}(1+R_t)]^{1/T} - 1$
Compounding	Ignores	Includes
Typical Use Case	Planning	Performance Reporting

Outline

Definition (Normal "Gaussian" Distribution)

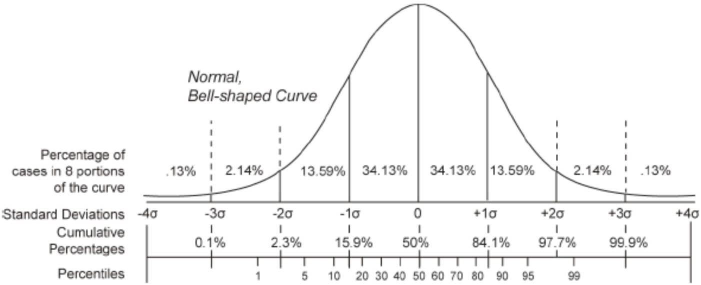
$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.

Definition (Normal "Gaussian" Distribution)

$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.



Definition (Normal "Gaussian" Distribution)

$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.

Definition (Normal "Gaussian" Distribution)

$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.

- $X \sim N(\mu, \sigma^2)$.

Definition (Normal "Gaussian" Distribution)

$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.

- $X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$, then $Z \equiv \frac{X-\mu}{\sigma} \sim N(0, 1)$.

Definition (Normal "Gaussian" Distribution)

$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.

- $X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$, then $Z \equiv \frac{X-\mu}{\sigma} \sim N(0, 1)$.
- $f(Z) \equiv \phi(Z)$, and $F(z) \equiv \Phi(z)$.

Definition (Normal "Gaussian" Distribution)

$$f(X) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

for $-\infty < X < \infty$.

- $X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$, then $Z \equiv \frac{X-\mu}{\sigma} \sim N(0, 1)$.
- $f(Z) \equiv \phi(Z)$, and $F(z) \equiv \Phi(z)$.
- If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.

Sum of Gaussian Random Variables

Definition (Sum of Normals)

If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$,and $Z = X + Y$, then

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$$

Sum of Gaussian Random Variables

Definition (Sum of Normals)

If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$,and $Z = X + Y$, then

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y)$$

Definition (Sum of Independent Normals)

If $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$,and $Z = X + Y$, then

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Modeling Prices, Not Returns

A common model of returns assumes $R_1, R_2, ...$ are iid $N(\mu, \sigma^2)$. Two problems arise:

1. Limited Liability ($P_t \geq 0 \rightarrow R_t \geq -1$) can be violated with Gaussian returns.
2. Even though R_t is Gaussian, multi-period returns $R_t(H) = \prod_{j=0}^{H-1}(1 + R_{t-j}) - 1$ are not.

Modeling Prices, Not Returns

A common model of returns assumes $R_1, R_2, \ldots$ are iid $N(\mu, \sigma^2)$. Two problems arise:

1. Limited Liability ($P_t \geq 0 \rightarrow R_t \geq -1$) can be violated with Gaussian returns.
2. Even though R_t is Gaussian, multi-period returns $R_t(H) = \prod_{j=0}^{H-1}(1 + R_{t-j}) - 1$ are not.

Solution:

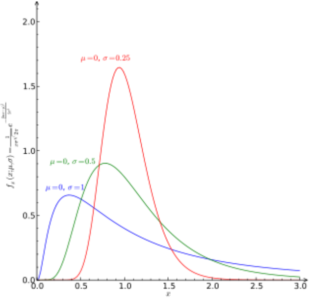
Assume $(1 + R) \sim LogNormal$

Implies that $r \equiv ln(1 + R) \sim N(\mu, \sigma^2)$

Definition (Log Normal Distribution)

$$f(X) = \frac{1}{X(2\pi)^{1/2}\sigma} exp\left[-\frac{1}{2}\left(\frac{ln(X)-\mu}{\sigma}\right)^2\right]$$

for $0 < X < \infty$. If $ln(X) \sim N$, then $X \sim$ log Normal. Especially useful for prices, rather than returns.



Modeling Prices, Not Returns

Resolving Limited Liability

Recall $r_t = ln(1 + R_t)$. Notice that $exp(r_t) = exp(ln(1 + R_t)) = 1 + R_t$. Recall that the exponential function has positive range, implying that $(1 + R_t) \geq 0 \rightarrow R_t \geq -1 \rightarrow P_t \geq 0$.

Resolving the lack of Multi-Period Gaussianity

$$
\begin{aligned}
1 + R_t(H) &= (1 + R_t)(1 + R_{t-1}).... \\
&= exp(r_t)exp(r_{t-1})... \\
&= exp(r_t + r_{t+1} + ...) \\
ln(1 + R_t) &= r_t + r_{t-1} + ...
\end{aligned}
$$

Recall that the **sum** of Gaussian r.v.'s is Gaussian.

Definition (Random Walk)

Let $Z_1, Z_2, ...$ be iid with mean μ and variance σ^2. Let P_0 be an arbitrary starting price. Then we can define the price process as a random walk if

$$P_t = P_0 + Z_1 + ...Z_t \; \forall \; t$$

Definition (Geometric Random Walk)

Recall $\frac{P_t}{P_{t-H}} = 1 + R_t(H) = exp(r_t + ... + r_{t-H+1})$, which implies

$$P_t = P_0 exp(r_t + ... + r_1)$$

If $r_1, r_2, ...$ are iid $N(\mu, \sigma^2)$ then we call the price process a lognormal geometric random walk with parameters (μ, σ^2).

Definition (Chi-Square Distribution)

Let $Z_i, i = 1, 2, ..., n$ be independent, identically distributed $N(0, 1)$. If,

$$X = \sum_{i=1}^{n} Z_i^2$$

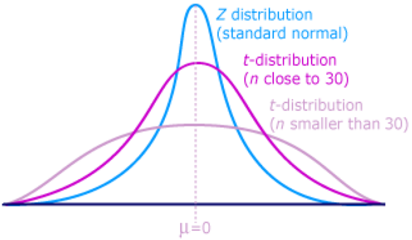
then $X \sim \chi_n^2$, where n corresponds to the degrees of freedom.

Definition (t-Distribution)

Let $Z \sim N(0,1)$ and $X \sim \chi^2_n$. Assume Z and X are independent. If,

$$T = \frac{Z}{\sqrt{X/n}}$$

then $T \sim t_n$, where N corresponds to the degrees of freedom.



Definition (F-Distribution)

Let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$. Assume X_1 and X_2 are independent. If

$$F = \frac{X_1/k_1}{X_2/k_2}$$

then $F \sim F_{k_1,k_2}$ where (k_1, k_2) are the degrees of freedom.

Outline

1. Compute Returns
2. Describe Returns
3. Inference

Reminder of Goal

- Hypothesis testing can help determine STATISTICAL significance
- Is the return $= 0$? Is the beta of this asset greater than 1? Is the standard deviation of cash flows for project A equal to that of project B?
- Relying exclusively on sample statistics may be misleading
- Variability in the sample data will influence our confidence that the the sample statistics match the true population measures
- Hypothesis testing is a way to account for this uncertainty
- This tool is broad, so in the following we will use a generic Y variable to be our object of interest

Outline

Anatomy of a Hypothesis Test

The null hypothesis (H_0) and the alternative hypothesis (H_1) of the one-sample t-test can be expressed as

$$H_0 : E[Y] = \mu_0$$
$$H_1 : E[Y] \neq \mu_0$$

H_0 – the population expectation $E[Y]$ is equal to the proposed (aka hypothesized) value μ_0
H_1 – the population expectation $E[Y]$ is NOT equal to the proposed (aka hypothesized) value μ_0

Anatomy of a Hypothesis Test

The alternative hypothesis $E[Y] \neq \mu_0$ contains two cases:

$$E[Y] > \mu_0 \text{ and } E[Y] < \mu_0$$

Hence, when cast this way, H_0 and H_1 form a two-sided test.

If the alternative hypothesis is cast as either of the inequalities above, we have a one-sided test.

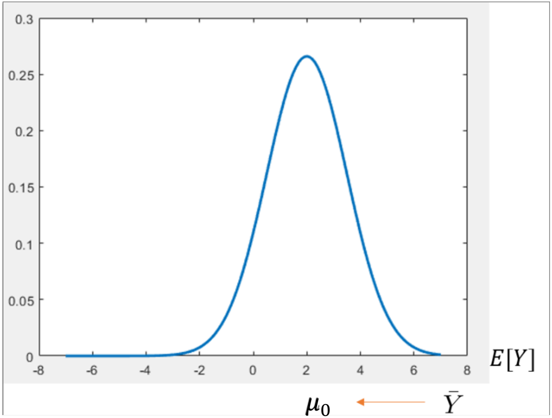
How do we choose Null and Alternative?

H_0 — Null Hypothesis is usually considered as established consensus. What is already believed. Least costly. That which requires no further action. Assumed to be true until "proven" otherwise.

H_1 — Alternative Hypothesis is something different than consensus. What you believe might be true, but is different from others. Not the base case. If true, leads to costly action.

Constructing a Test

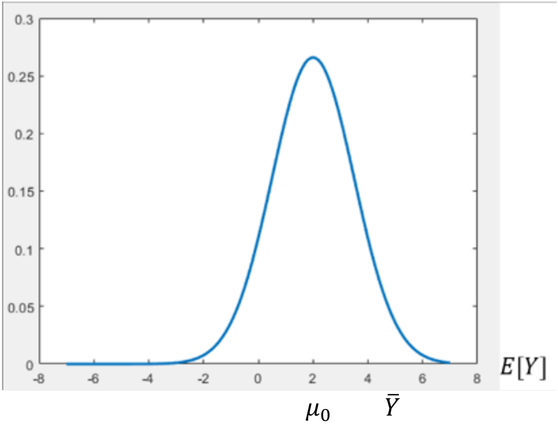
Suppose we are using $\bar{Y}$ as an estimator for $E[Y]$ and want to test $H_0 : E[Y] = \mu_0$.



Is $\bar{Y}$ close enough to μ_0?

Constructing a Test

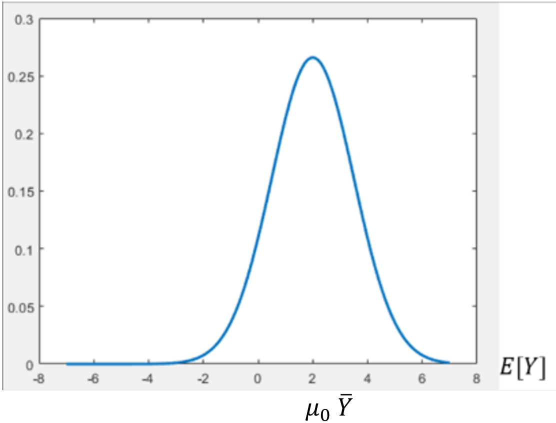
Suppose we are using $\bar{Y}$ as an estimator for $E[Y]$ and want to test $H_0 : E[Y] = \mu_0$.



Is $\bar{Y}$ close enough to μ_0? - How about now?

Constructing a Test

Suppose we are using $\bar{Y}$ as an estimator for $E[Y]$ and want to test $H_0 : E[Y] = \mu_0$.



How close do μ_0 and $\bar{Y}$ have to be to say they are equal?

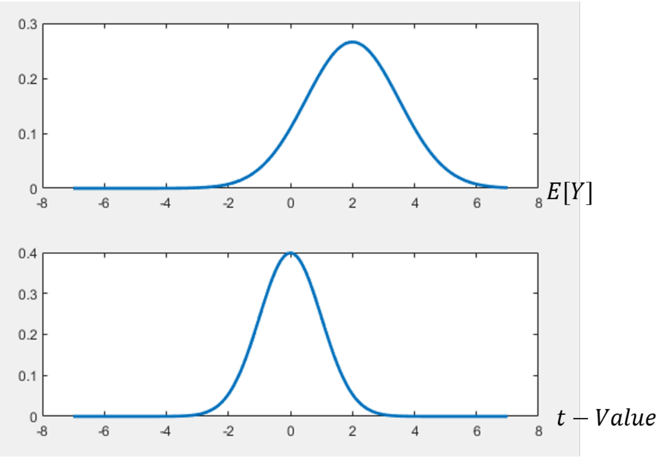
Constructing a Test

A hypothesis testing procedure can answer that question.

1. Rescale (standardize) the data to generate a test statistic
2. Establish critical values
3. Compare the critical values to the observed test statistic

Constructing a Test
Rescaling

The first thing we do is to rescale (standardize) the data.



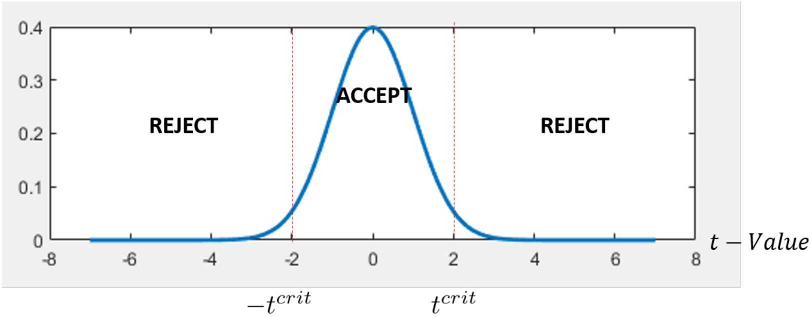
One Sample t-test of the Mean

$H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$

$$t^{stat} = \frac{\bar{Y} - \mu_0}{s_Y/\sqrt{N}}$$

where

- t^{stat} = test statistic
- μ_0 = Proposed constant for the population expectation
- $\bar{Y}$ = Sample mean
- N = Sample size (i.e. number of observations)
- s_Y = Sample standard deviation

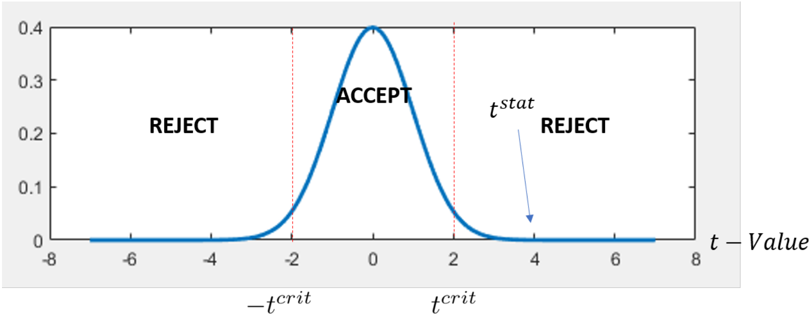


where t^{crit} is the "critical" value.

Constructing a Test

Compare Test Statistic to Critical Value

Reject H_0 if $|t^{stat}| > t^{crit}$
Fail to Reject H_0 otherwise.



Fail to Reject vs Accept

Why do we say "Fail to Reject" rather than "Accept"?

Fail to Reject vs Accept

Why do we say "Fail to Reject" rather than "Accept"?

- Suppose our proposed value $\mu_0 = 5$. Suppose our estimate $\bar{Y} = 4.95$.

Fail to Reject vs Accept

Why do we say "Fail to Reject" rather than "Accept"?

- Suppose our proposed value $\mu_0 = 5$. Suppose our estimate $\bar{Y} = 4.95$.
- Consider the test $H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$

Fail to Reject vs Accept

Why do we say “Fail to Reject” rather than “Accept”?

- Suppose our proposed value $\mu_0 = 5$. Suppose our estimate $\bar{Y} = 4.95$.
- Consider the test $H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$
- Suppose that given the sampling error, our data is consistent with the null.

Fail to Reject vs Accept

Why do we say "Fail to Reject" rather than "Accept"?

- Suppose our proposed value $\mu_0 = 5$. Suppose our estimate $\bar{Y} = 4.95$.
- Consider the test $H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$
- Suppose that given the sampling error, our data is consistent with the null.
- We are NOT saying that 4.95=5.

Fail to Reject vs Accept

Why do we say "Fail to Reject" rather than "Accept"?

- Suppose our proposed value $\mu_0 = 5$. Suppose our estimate $\bar{Y} = 4.95$.
- Consider the test $H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$
- Suppose that given the sampling error, our data is consistent with the null.
- We are NOT saying that 4.95=5.
- We are saying that 4.95 is statistically close enough to 5.

Fail to Reject vs Accept

Why do we say "Fail to Reject" rather than "Accept"?

- Suppose our proposed value $\mu_0 = 5$. Suppose our estimate $\bar{Y} = 4.95$.
- Consider the test $H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$
- Suppose that given the sampling error, our data is consistent with the null.
- We are NOT saying that 4.95=5.
- We are saying that 4.95 is statistically close enough to 5.
- Hence, we are NOT "Accepting" the null. We "Fail to Reject".

How Set Critical Values?

Where do we draw the critical values?
Depends on what types of mistakes we are willing to permit.

How Set Critical Values?

Where do we draw the critical values?
Depends on what types of mistakes we are willing to permit.

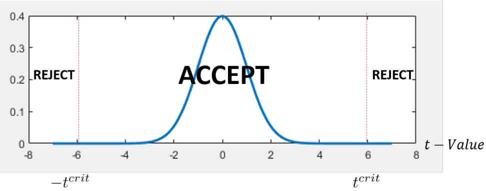
	H_0 True	H_0 False
"Accept" H_0		Type 2 Error (false -)
"Reject" H_0	Type 1 Error (false $+$) α, size	Power

How Set Critical Values?

Suppose you want to reduce the Type 1 error. How set critical values? What is the "side effect"?

How Set Critical Values?

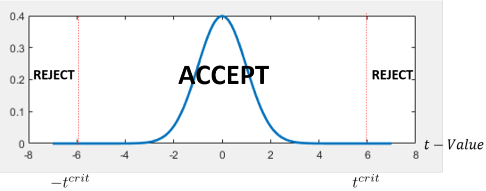
Suppose you want to reduce the Type 1 error. How set critical values? What is the "side effect"?



Type 2 error is higher
More likely that we accept the false H_0
Pr(Accept $H_0|H_0$ False) is High

How Set Critical Values?

Suppose you want to reduce the Type 1 error. How set critical values? What is the "side effect"?

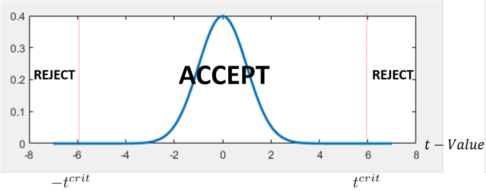


Type 2 error is higher
More likely that we accept the false H_0
Pr(Accept $H_0 | H_0$ False) is High

Suppose you want to reduce the Type 2 error. How set critical values? What is the "side effect"?

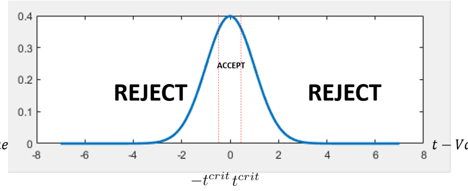
How Set Critical Values?

Suppose you want to reduce the Type 1 error. How set critical values? What is the "side effect"?



Type 2 error is higher
More likely that we accept the false H_0
$Pr(\text{Accept } H_0 | H_0 \text{ False})$ is High

Suppose you want to reduce the Type 2 error. How set critical values? What is the "side effect"?



Type 1 error is higher
More likely that we reject the true H_0
$Pr(\text{Reject } H_0 | H_0 \text{ True})$ is High

How Set Critical Values?

Why do Judges say “Not Guilty” instead of “Innocent”?

How Set Critical Values?

Why do Judges say "Not Guilty" instead of "Innocent"?

	H_0 True	H_0 False
"Accept" H_0		Type 2 Error (false -) *Let Guilty go Free*
"Reject" H_0	Type 1 Error *Send Innocent to Jail*	Power

How Set Critical Values?

Why do Judges say “Not Guilty” instead of “Innocent”?

	H_0 True	H_0 False
“Accept” H_0		Type 2 Error (false -) *Let Guilty go Free*
“Reject” H_0	Type 1 Error *Send Innocent to Jail*	Power

U.S. Legal convention is to minimize Type 1 Error. That is viewed as the more important mistake to avoid.

How Set Critical Values?

Why do Judges say "Not Guilty" instead of "Innocent"?

	H_0 True	H_0 False
"Accept" H_0		Type 2 Error (false -) *Let Guilty go Free*
"Reject" H_0	Type 1 Error *Send Innocent to Jail*	Power

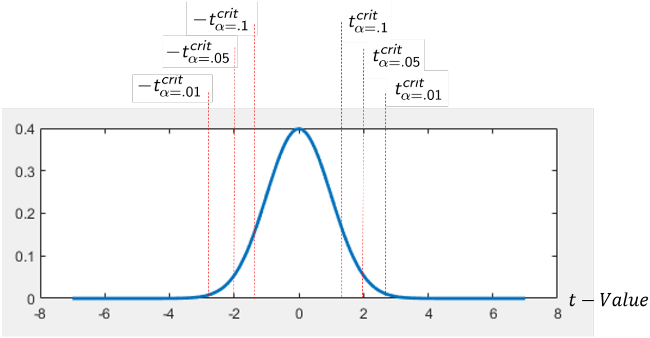
U.S. Legal convention is to minimize Type 1 Error. That is viewed as the more important mistake to avoid. H_0 : Innocent; H_A : Guilty.

How Set Critical Values?

Critical values typically chosen to set Type 1 Error $\approx \{1\%, 5\%, 10\%\}$

How Set Critical Values?

Critical values typically chosen to set Type 1 Error $\approx \{1\%, 5\%, 10\%\}$



As we permit smaller Type 1 Errors, the critical value grows (in absolute terms), implying that we are less likely to reject a null hypothesis (i.e. Type 2 error rises)

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) = N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) $=$ N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

As df rises, the critical value falls, implying more likely to reject.

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) = N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

As df rises, the critical value falls, implying more likely to reject.

Small sample sizes imply df are

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) = N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

As df rises, the critical value falls, implying more likely to reject.

Small sample sizes imply df are low, which means critical values are

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) = N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

As df rises, the critical value falls, implying more likely to reject.

Small sample sizes imply df are low, which means critical values are higher, so the likelihood to reject

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) = N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

As df rises, the critical value falls, implying more likely to reject.

Small sample sizes imply df are low, which means critical values are higher, so the likelihood to reject falls, and Type 2 Error

How Set Critical Values?

Critical values are also as function of the sample size (N).

The sample size influences the *degrees of freedom*.

- Degrees of Freedom: df (or dof) = N-1
- As the df grows, the distribution of the distribution of the t-statistic approaches Normal distribution

As df rises, the critical value falls, implying more likely to reject.

Small sample sizes imply df are low, which means critical values are higher, so the likelihood to reject falls, and Type 2 Error rises.

Outline

3. Inference
    - Anatomy of a Hyp. Test
    - Evaluating a Hyp. Test
    - Common Tests

Evaluating the Test

Definition (p-Value)

A p-Value is the largest significance level at which we could carry out the test and still fail to reject the null hypothesis

$$
\begin{aligned}
p = \quad & 2Pr(t > t^{stat}) \\
& 2[1 - F(t^{stat})]
\end{aligned}
$$

Example: p-Val=.03 means 3% chance of getting a $t - value$ "greater" than the t^{stat}.
Slightly more precisely, there is a 1.5% chance of getting a t-value greater than the t^{stat} and 1.5% chance of getting a t-value less than the $-t^{stat}$.

Evaluating the Test

One Sample t-test of the Mean

$H_0 : E[Y] = \mu_0$ vs $H_1 : E[Y] \neq \mu_0$

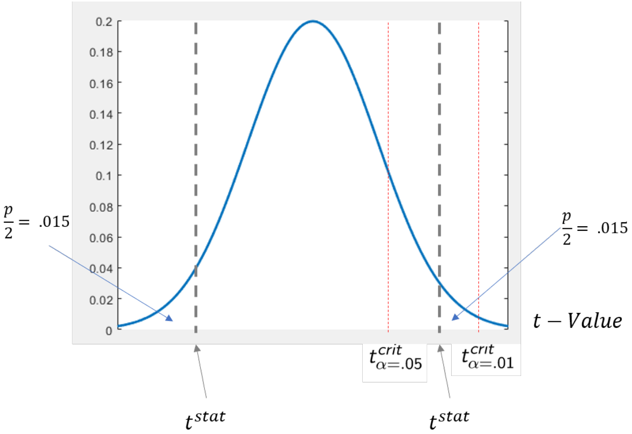
$$t^{stat} = \frac{\bar{Y} - \mu_0}{s_Y/\sqrt{N}}$$

where

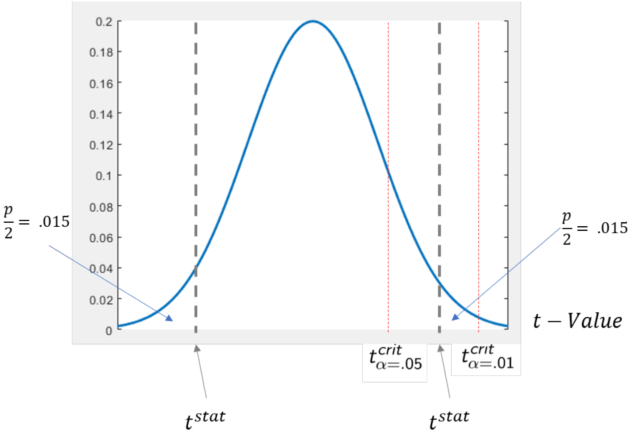
- t^{stat} = test statistic
- μ_0 = Proposed constant for the population expectation
- $\bar{Y}$ = Sample mean
- N = Sample size (i.e. number of observations)
- s_Y = Sample standard deviation

Reject H_0 if $|t^{stat}| > t^{crit}$ or if $p - Val < \alpha$

Evaluating the Test

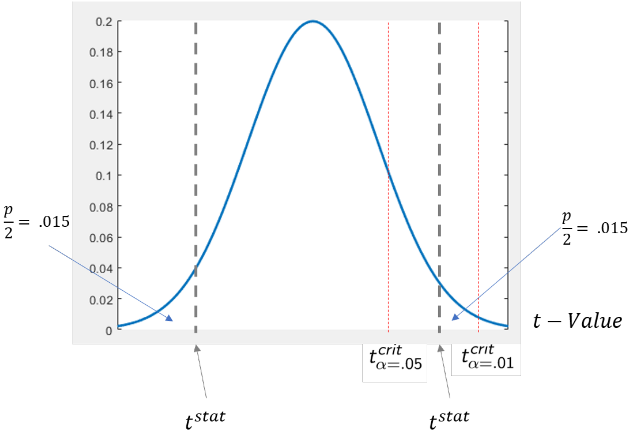


Evaluating the Test



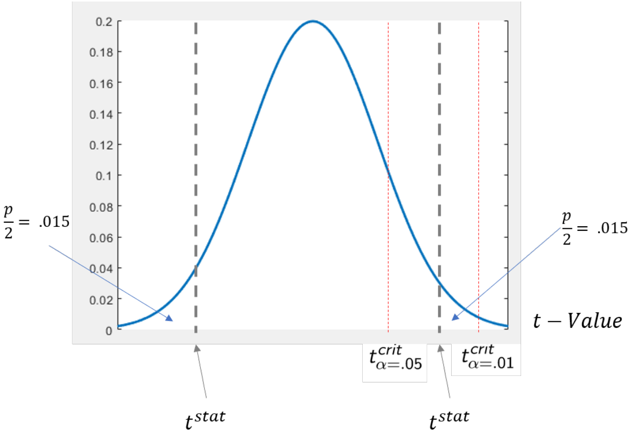
- Since $t^{stat} > t^{crit}_{.05}$ we reject at the 5% level

Evaluating the Test



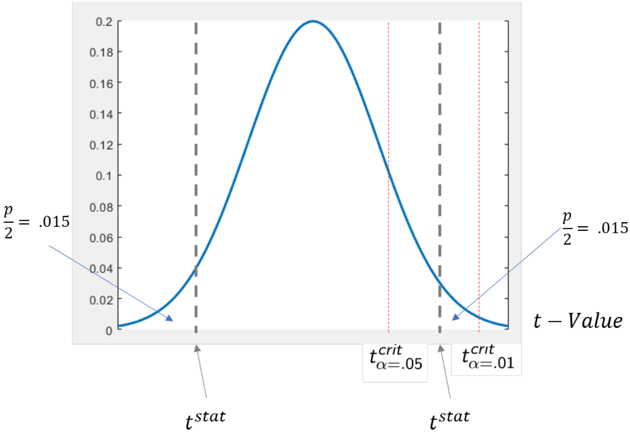
- Since $t^{stat} > t^{crit}_{.05}$ we reject at the 5% level
- Since $t^{stat} < t^{crit}_{.01}$ we fail to reject at the 1% level

Evaluating the Test



- Since $t^{stat} > t^{crit}_{.05}$ we reject at the 5% level
- Since $t^{stat} < t^{crit}_{.01}$ we fail to reject at the 1% level
- Rule of Thumb: Reject if $t^{stat} > 2$

Evaluating the Test



- Since $t^{stat} > t^{crit}_{.05}$ we reject at the 5% level
- Since $t^{stat} < t^{crit}_{.01}$ we fail to reject at the 1% level
- Rule of Thumb: Reject if $t^{stat} > 2$
- Since $p = .015 * 2 = .03$, we fail to reject the null at $\alpha < .03$

Outline

3 Inference

- Anatomy of a Hyp. Test
- Evaluating a Hyp. Test
- Common Tests

Testing Equality of Two Means

Myriad tests are possible (means, variances, etc..). A common application you will face is testing the equality of two averages.

$$H_0: \quad E[Y_1] - E[Y_2] = \mu_0$$
$$H_1: \quad E[Y_1] - E[Y_2] \neq \mu_0$$

By setting $\mu_0 = 0$ we can test equality of the two series' expected values

Testing Equality of Two Means

Let's use the sample averages to determine if the expected values are equal.
$H_0 : E[Y_1] = E[Y_2]$ vs $H_1 : E[Y_1] \neq E[Y_2]$

$$t^{stat} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s^2_{Y_1}}{N_1} + \frac{s^2_{Y_2}}{N_2}}}$$

where

- $\bar{Y}_i$ = Sample mean (i=1,2)
- N_i = Sample size (i=1,2)
- $s^2_{Y_i}$ = Sample Variance (i=1,2)

Typical implementation is via Welch ('47, '51)

Testing Equality of Two Proportions

Suppose you want to determine if two proportions are equal. Then we can establish $H_0 : p_A = p_B$ vs $H_1 : p_A \neq p_B$ via

$$z = \frac{p_A - p_B}{\sqrt{p(1-p)/N_A + p(1-p)/N_B}}$$

where

- $\tilde{p}_i$ as the true proportion for group i and p_i as its sample counterpart
- N_i as the sample size of group i
- $p = (p_a N_A + p_B N_B)/(N_A + N_B)$

This test statistic follows a standard normal, but other implementations are available (e.g. Chi-square tests).