Econometrics: Inferential Statistics Handout

Ryan Safner

Fall 2017

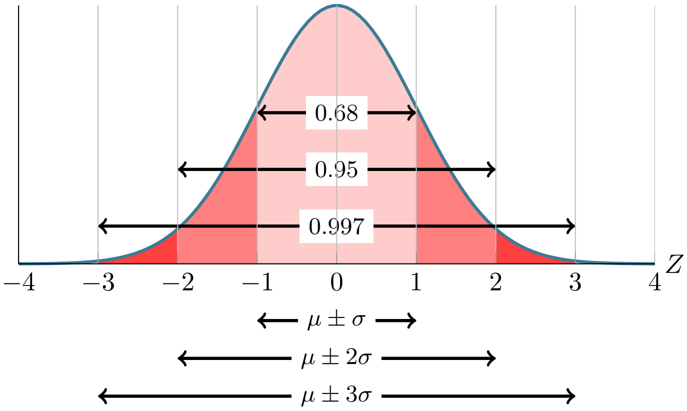
1 Normal Distribution

- Normal distribution is a continuous distribution of a random variable

$$X \sim N(\mu, \sigma)$$

  - Mean μ
  - Standard deviation σ

- 68-95-99.7% empirical rule:
  - $P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.68$
  - $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$
  - $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997$



  - Can find the probability of X being a certain range.
    - $P(a \leq X \leq b)$ can be calculated using `normalcdf(a, b,` μ`,` σ`)`
    - For a right or left bound that is not a value (e.g. $P(X > a)$ or $P(b < X)$), use positive or negative infinity with `1E99`)
  - Can find the value of X that is a certain percentile
    - `InvNorm(`ϕ, μ, σ`)` where ϕ is the area under the pdf to the left of the unknown value (i.e. the cdf)
  - Standard Normal Distribution: $X \sim N(0, 1)$
    - Standardize a random variable by calculating it's Z-score:

$$Z = \frac{X_i - \mu}{\sigma}$$

    - Z is the number of standard deviations above $(+)$ or below $(-)$ the mean a value is
  - We use Z-scores and the standard normal to find the probability between a range under the curve
    - With standard normal distribution, only need the two boundaries: `normalcdf(LB, RB)`
  - Can find the Z-score for a certain percentile analogous to finding the X value for the percentile
    - InvNorm(ϕ) where ϕ is the area under the pdf to the left of the unknown Z-score (i.e. the cdf)

2 Central Limit Theorem

- Inferential statistics
  - There are unknown *parameters* that describe a *population* distribution that we want to know
  - We use *statistics* generated from a *sample* to *estimate* the population parameters
- Sampling Distributions
  - Conducting multiple samples and generating statistics (e.g. the sample mean, $\bar{X}$) will naturally yield slightly different values for the statistics, there is *sampling variability*
  - The statistics (e.g. $\bar{X}$) themselves become random variables with their own distribution, called the *sampling distribution* of the sample statistic
    - Sampling distribution has a mean, $E(\bar{X}) = \mu_X$ (true population mean), and standard error $\sigma_{\bar{X}}$
- Central Limit Theorem
  - With large enough sample size ($n \geq 30$), the sampling distribution of a sample statistic is approximately normal
  - If samples are i.i.d. (independently and identically distributed if they are drawn from the same population randomly and then replaced) we don't even need to know the population distribution to assume normality
- Sampling distribution of the sample mean ($\bar{X}$):

$$\bar{X} \sim \left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

  - Standard error of the sample mean: $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$
  - Note: σ_X is the population standard deviation of X vs. $\sigma_{\bar{X}}$ is the standard deviation of the sample mean
  - Note: We need to know the population standard deviation σ_X
- We can find probabilities that the sample mean is within a certain range using the standard normal distribution using `normalcdf(LB, RB,` $\mu_X, \frac{\sigma_X}{\sqrt{n}}$`)`

- Sampling distribution of sample sums ($\sum X$) (adding up all values in sample)

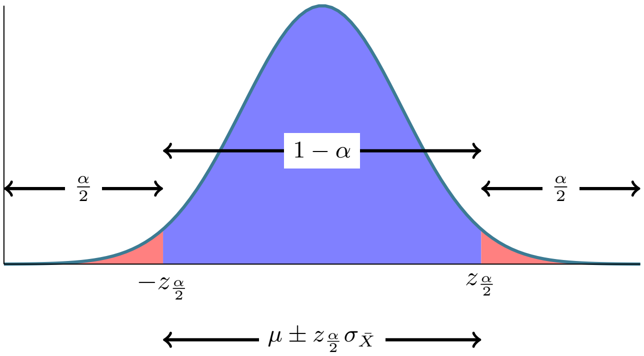
$$\bar{X} \sim \left(n\mu_X, \sigma_X\sqrt{n}\right)$$

3 Confidence Intervals

- A confidence interval describes the range of estimates for a population parameter of the form:

$$(\text{point estimate} - \text{margin of error}, \text{point estimate} + \text{margin of error})$$

- Our confidence level is $1 - \alpha$
  - α: significance level, the probability the true population parameter is *not* contained within our confidence interval
  - Typical confidence levels: 90%, 95%, 99%, especially 95%
- A confidence interval tells us that if we were to conduct many samples, $(1 - \alpha)\%$ would contain the true population parameter within the estimated range of values
- We need to know the *critical value* of $Z_{\frac{\alpha}{2}}$ on the pdf that puts $(1 - \alpha)$ probability between $\pm Z_{\frac{\alpha}{2}}$ and $\frac{\alpha}{2}$ probability in each of the tails beyond $\pm Z_{\frac{\alpha}{2}}$
  - Common values:
    - For $\alpha = 0.10$: 1.65
    - For $\alpha = 0.05$: 1.96
    - For $\alpha = 0.99$: 2.58
  - These values can be found with `InvNorm(`$[1 - \frac{\alpha}{2}]$`)`



- **Confidence intervals for *means*:**
  - **If we know the population standard deviation σ_X AND $n \geq 30$**
    - We can use the standard normal distribution
    - The *margin of error (MOE)* is the critical value of Z times the standard error of the estimate:

$$MOE = Z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}$$

    - So the full confidence interval is:

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma_X}{\sqrt{n}}$$

  - **If we don't know the population standard deviation σ_X OR $n < 30$**
    - We need to use the Student's t-distribution with $n - 1$ degrees of freedom (df)
    - We use sample standard deviation (s_X) to estimate population standard deviation (σ_X)

        - We calculate the t-score, analogous to Z-scores:

$$t = \frac{\bar{X} - \mu}{(\frac{s}{\sqrt{n}})}$$

        - We need to find the critical value of t that puts $\frac{\alpha}{2}$ in each tail, this is different for each t distribution based on *df*
            - Some calculators can do `InvT([1 −` $\frac{\alpha}{2}$`])`, otherwise need a t-table to look up critical values
    - The *margin of error (MOE)* is the critical value of t times the standard error of the estimate:

$$MOE = t_{\frac{\alpha}{2}} \frac{s_X}{\sqrt{n}}$$

    - So the full confidence interval is:

$$\bar{X} \pm t_{\frac{\alpha}{2}} \frac{s_X}{\sqrt{n}}$$

- **Confidence intervals for *proportions***
    - We need to ensure the underlying distribution is binomial
        - Recall it must be a series of n identical and independent trials, with each trial resulting in either "success" with probability p or "failure" with probability $(1 - p)$; X is the number of successes
        - $E(X) = np$
        - $\sigma_X = \sqrt{np(1-p)}$
    - We estimate the population proportion (of successes) by calculating a sample proportion $\hat{p}$:

$$\hat{p} = \frac{X}{n}$$

    - Under the central limit theorem, with large enough $np > 5$, normal distributions approximate the binomial distribution of X:

$$X \sim N(np, \sqrt{np(1-p)})$$

    - The sampling distribution of the sample proportion is normally distributed:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

        - Mean: p (true population proportion)
        - Standard deviation: $\sqrt{\frac{p(1-p)}{n}}$
    - The *margin of error (MOE)* is the critical value of Z times the standard error of the estimate:

$$MOE = Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

    - So the full confidence interval is:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

    - Can confirm calculations with calculator using `STAT` → `TESTS` and either `ZInterval` for means knowing σ and $n > 30$, `TInterval` for means without knowing σ or $n < 30$, or `1-PropZInterval` for proportions

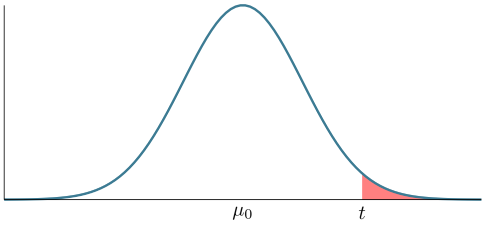
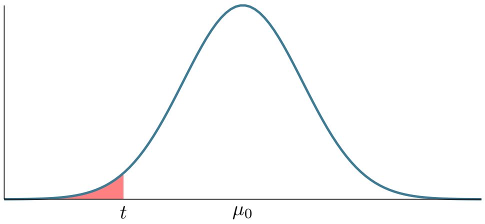
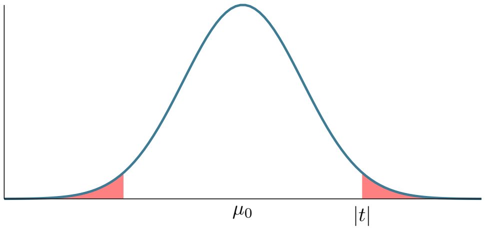
4 Hypothesis Testing

- We want to test whether a population parameter is likely to be a hypothesized value (or range) vs. an alternate proposed value (or range)
  - Null hypothesis, H_0: population parameter is some value or range
  - Alternate hypothesis, H_A that must mathematically contradict H_0
    - *Two-sided alternative* (e.g. that the parameter $\neq H_0$ value)
    - *One-sided alternative* (e.g. that the parameter $>$ OR $< H_0$ value)
  - Always test whether or not our sample statistics provide sufficient evidence to reject H_0 in favor of H_A
- Sample statistics might differ from true population parameters
  - EITHER due to the fact that our hypothesized population parameter is false, OR that it is actually true but our sample gives us a different estimate due to natural sampling variability (we do not know which is correct)
  - Type I error (false positive): rejecting H_0 when H_0 is in fact true
    - $\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true})$
    - α is the significance level (e.g. 0.01, 0.05, 0.10)
  - Type II error (false negative): failing to reject H_0 when H_0 is in fact false
    - $\beta = P(\text{Type II error}) = P(\text{Don't reject } H_0 | H_0 \text{ is false})$
    - $1 - \beta =$ *Power* of the test

	H_0 is True	H_0 is False
Reject H_0	Type I Error False Positive α	Correct Outcome True Positive $1-\beta$
Don't Reject H_0	Correct Outcome True Negative $1-\alpha$	Type II Error False Negative β

  - p-value is the probability that, if the null hypothesis were true, we would obtain a result at least as extreme as the one in our sample
    - If $p < \alpha$, the finding is "statistically significant," and we have sufficient evidence to reject H_0
    - If $p > \alpha$, the finding is not statistically significant, and we do not have sufficient evidence to reject H_0
- All tests take the form of comparing the value of a test statistic to a critical value of a distribution (e.g. Z or t) or computing the p-value from that test statistic value

$$\text{test statistic} = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error of the estimate}}$$

Alternative	p-value	PDF
$H_a : \mu > \mu_0$	$P(T \geq t)$	
$H_a : \mu < \mu_0$	$P(T \leq t)$	
$H_a : \mu \neq \mu_0$	$2[P(T \geq \lvert t \rvert)]$	

- **Hypothesis testing of a sample mean**

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

$$H_2 : \mu > \mu_0$$

$$H_3 : \mu < \mu_0$$

  - **If we know population standard deviation σ and $n > 30$**
    - We run a Z-test

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

    - If computed test statistic $z \geq z*$ (depending on alternative hypothesis and α-level), we are in the rejection region and can reject H_0
    - Can compute the actual p-value for $P(Z \geq z)$ using `normalcdf`
  - **If we don't know population standard deviation σ OR $n < 30$**
    - We run a t-test on a t-distribution with $n - 1$ degrees of freedom

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

    - If computed test statistic $z \geq z*$ (depending on alternative hypothesis and α-level), we are in the rejection region and can reject H_0
    - Can compute the actual p-value for $P(Z \geq z)$ using `normalcdf`

- **Hypothesis testing of a sample proportion**

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

$$H_2 : p > p_0$$

$$H_3 : p < p_0$$

  - We run a Z-test

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

  - If computed test statistic $z \geq z*$ (depending on alternative hypothesis and α-level), we are in the rejection region and can reject H_0
  - Can compute the actual p-value for $P(Z \geq z)$ using `normalcdf`

- **Hypothesis testing of a difference in sample means (e.g. 2 groups, X and Y)**

$$H_0 : \mu_X - \mu_Y = d_0$$

$$H_1 : \mu_X - \mu_Y \neq d_0$$

$$H_2 : \mu_X - \mu_Y > d_0$$

$$H_3 : \mu_X - \mu_Y < d_0$$

  - **If we know population standard deviation σ and $n > 30$ for both X and Y, we run a Z-test**

$$z = \frac{(\bar{X} - \bar{Y}) - d_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

    - **If we don't know population standard deviation σ OR $n < 30$ for X or Y, we run a t-test**

      $$t = \frac{(\bar{X} - \bar{Y}) - d_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

    - We run a t-test on a t-distribution with degrees of freedom according to a very complex formula that your calculator/software can calculate
    - If computed test statistic z (or t)$\geq$ $z*$ (or $t*$) (depending on alternative hypothesis and α-level), we are in the rejection region and can reject H_0
    - Can compute the actual p-value using `normalcdf`
  - Can confirm calculations with calculator using `STAT` $\rightarrow$ `TESTS` and either `Z-Test` for means knowing σ and $n > 30$, `T-Test` for means without knowing σ or $n < 30$, `1-PropZTest` for proportions, `2-SampZTest` for difference in means knowing σ and $n > 30$, `2-SampTTest` for means without knowing σ or $n < 30$