# ECOM20001
# Econometrics 1

## Lecture Note 3
## Statistics

A/Prof David Byrne
Department of Economics
University of Melbourne

Stock and Watson: Chapter 3

# Summary of Key Concepts

- Estimators and sample averages
- Hypothesis tests of means and p-values
- Sample variance, sample standard deviation, standard error
- One-sided alternatives
- Confidence intervals
- Comparing means from different populations
- t-statistics with small sample sizes
- Scatterplots, sample covariance and correlation

# Introduction

- In lecture note 2 we discussed probability, which is the study of distributions of random variables
  - could involve 1 variable: what is the mean income level?
  - could involve 2 or more variables: what are mean income levels for males and females?; what is the correlation between the number of pokies per person and average income in an LGA?
- This lecture note reviews statistics, which is the study of testing hypotheses about distributions of random variables
- Key insight of statistics: we can learn about the population distribution from a random sample from that population
- Will continue to focus on the pokies dataset and sample averages means
- Some of this material may be review, other parts may not be. All of it is building blocks for regression analysis.

# Estimators

- Throughout we will work with an iid random sample $Y_1, Y_2, \ldots, Y_n$ from the population with $n$ observations
- Let's first focus on the sample average $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, an estimator of underlying population mean $\mu_Y$
- Two key properties of $\bar{Y}$ as an estimator of $\mu_Y$
  1. Unbiasedness: $E(\bar{Y}) = \mu_Y$
     → in expectation, the sample average gives us the right answer about the underlying population mean
  2. Consistency: as $n \uparrow$, $var(\bar{Y}) = \frac{\sigma_Y^2}{n} \downarrow$
     → as the sample size gets bigger, the variance of our estimator $\bar{Y}$ falls implying that the sample average is more likely to give the right answer about the underlying population mean

Note: skipping $\bar{Y}$ is BLUE and $\bar{Y}$ as a least square estimator of $\mu_Y$ on p. 114-115 in the text. Will return to this later with regression analysis.

# Hypothesis Testing with the Population Mean

- We can use a random sample and the sample average $\bar{Y}$, to test hypotheses about the underlying population mean $\mu_Y$
- The null hypothesis is the claim that we are testing

$$H_0 : E(Y) = \mu_{Y,0}$$

- The alternative hypothesis is the claim the null is being tested against

$$H_1 : E(Y) \neq \mu_{Y,0}$$

# Hypothesis Testing with the Population Mean Example

- Politician says "we don't have a gambling problem - there's only 1 pokie per 1000 people on average!"
  - $H_0 : E(\texttt{Negms\_1000}) = 1; \quad H_1 : E(\texttt{Negms\_1000}) \neq 1$
- Hypotheses involving $\neq$ are commonly called two-sided alternative hypothesis
  - they hypothesis can be wrong in two ways: either the true value of the mean relative to the hypothesized is $\mu_Y > \mu_{Y,0}$ or $\mu_Y < \mu_{Y,0}$

# p-values

- There are two reasons why $\bar{Y}$ and $\mu_{Y,0}$ might differ:
    1. The null is false: the true value of $\mu_Y$ does not equal $\mu_{Y,0}$
    2. Random sampling: the true value of $\mu_Y$ indeed equals $\mu_{Y,0}$, but the sample we randomly drew just happened to produce a $\bar{Y}$ that differs from $\mu_{Y,0}$ (say, by a lot)
- If we observe $\bar{Y} \neq \mu_{Y,0}$, the p-value, also known as the significance probability, helps us determine whether a false null or random sampling was factor that caused this difference

# p-values

- To construct a p-value, we first need to know what the probability distribution of $\bar{Y}$ is
- Recall from lecture note 2, from the Central Limit Theorem (CLT) we know that $\bar{Y}$ is distributed $N(\mu_Y, \sigma^2_{\bar{Y}})$ if sample size $n$ is large enough
  - $\mu_Y$ is the population mean
  - $\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$ is the standard deviation of $\bar{Y}$ where $\sigma_Y$ is the population standard deviation and $n$ is sample size

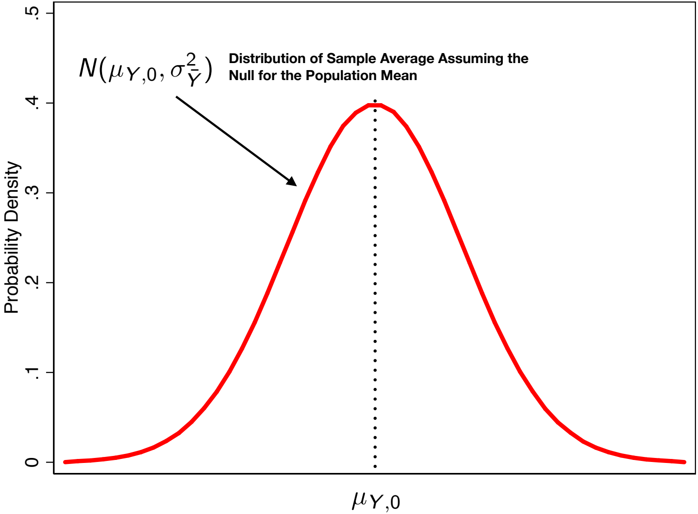
# p-values

- Given the CLT, and assuming the null hypothesis is correct (e.g., $\mu_Y = \mu_{Y,0}$), $\bar{Y}$ would be distributed:

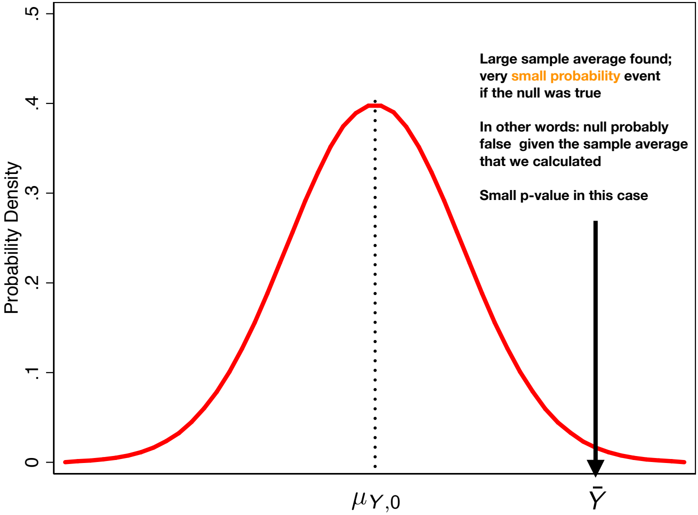
  $$N(\mu_{Y,0}, \sigma^2_{\bar{Y}})$$

  with $\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$
- Intuitively, p-values ask: what is the chance of obtaining a sample average at least as extreme as $\bar{Y}$ if the null was true and the underlying population mean was in fact $\mu_{Y,0}$?

# p-values Intuition Graphically



# p-values Intuition Graphically

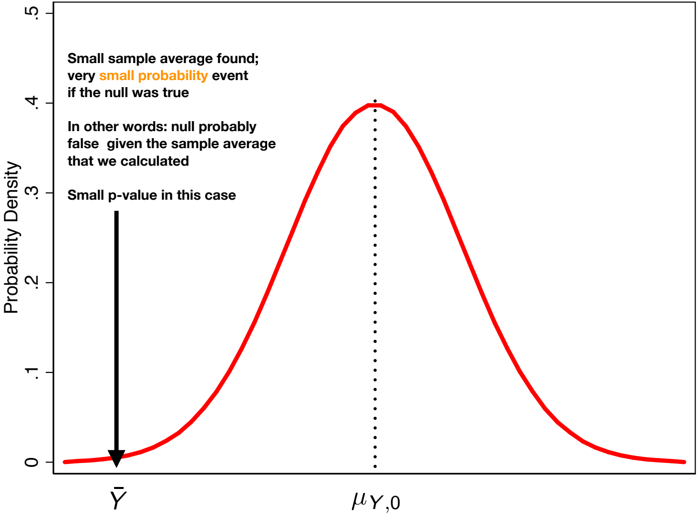


Large sample average found; very small probability event if the null was true

In other words: null probably false given the sample average that we calculated

Small p-value in this case

# p-values Intuition Graphically

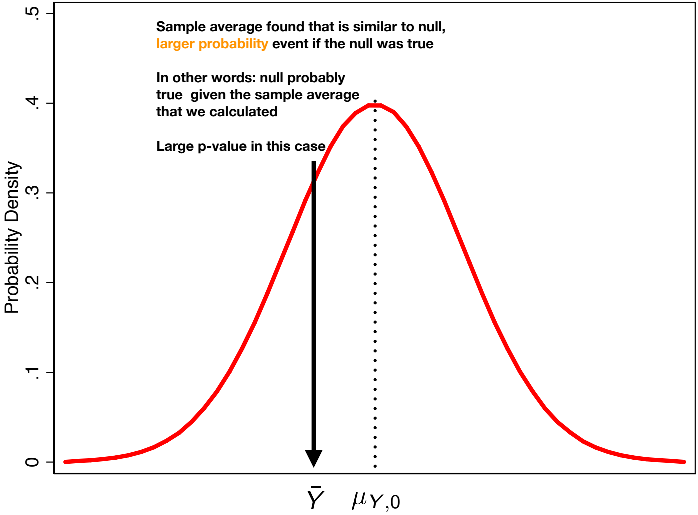


Small sample average found;
very small probability event
if the null was true

In other words: null probably
false given the sample average
that we calculated

Small p-value in this case

# p-values Intuition Graphically

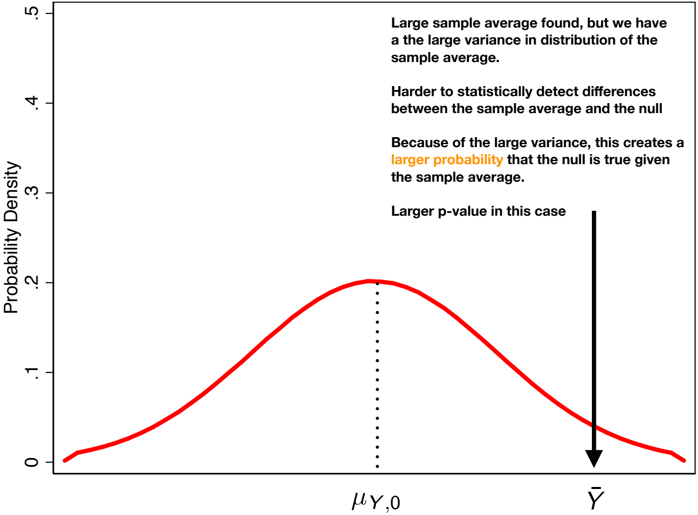


Sample average found that is similar to null, larger probability event if the null was true

In other words: null probably true  given the sample average that we calculated

Large p-value in this case

# p-values Intuition Graphically



Large sample average found, but we have a the large variance in distribution of the sample average.

Harder to statistically detect differences between the sample average and the null

Because of the large variance, this creates a larger probability that the null is true given the sample average.

Larger p-value in this case

# p-values example

- Consider the first example above where a large sample average was found: $\bar{Y} - \mu_{Y,0}$ is big and $\sigma_{\bar{Y}}$ is small
- What does this mean?
- $\bar{Y} - \mu_{Y,0}$ implies a big difference between the sample average $\bar{Y}$ and hypothesized population value $\mu_{Y,0}$
- At the same time $\sigma_{\bar{Y}}$ means we have a relatively precise estimate of $\bar{Y}$ (say, compared to the last example where there was larger variance)
- With a precise $\bar{Y}$ estimate that is far away from $\mu_{Y,0}$ should lead to be confident that $H_0 : \mu_Y = \mu_{Y,0}$ is probably not true
  - it would be unlikely to get a large sample average like $\sigma_{\bar{Y}}$ if $H_0$ were true

# p-values

- Rather than working with the $N(\mu_{Y,0}, \sigma^2_{\bar{Y}})$ distribution of the sample average $\bar{Y}$, we equivalently work with a standardized normal random variable $z$ (called a "z-score") in constructing p-values for a given null $H_0$:

  $$z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$$

  where $z$ is distributed $N(0, 1)$

# p-values

- $z$-score

$$z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$$

  where $z$ is distributed $N(0,1)$
- Remember, the basic idea is twofold:
  - false null: if $\bar{Y} - \mu_{Y,0}$ is big (causes $\uparrow z$), then $\mu_{Y,0}$ is more likely to be false
  - random sampling: if $\sigma_{\bar{Y}}$ is big (causes $\downarrow z$), then $\bar{Y}$ is a less precise estimate of $\mu_Y$, and any differences between $\bar{Y}$ and $\mu_{Y,0}$ are more likely driven by random sampling
- Competiting forces of "false null" and "random sampling" determine the probability that a null is true, given our random sample and the sample average we computed
  $\rightarrow$ big $z$-score means null is less likely to be true

# Computing p-values

- Suppose that $\bar{Y}^{act}$ is the actual sample average from our random sample, and assume for now that $\sigma_{\bar{Y}}$ is known (we will come back to this later).
- Then the corresponding z-score is

$$z^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$$

- p-values formally ask: what is the chance of obtaining a sample average at least as extreme as $\bar{Y}$ if the null was true?
  - Recall that $\bar{Y}$ can be extremely different in magnitude from $\mu_{Y,0}$ if $\bar{Y}$ is really big OR if $\bar{Y}$ is really small

# Computing p-values

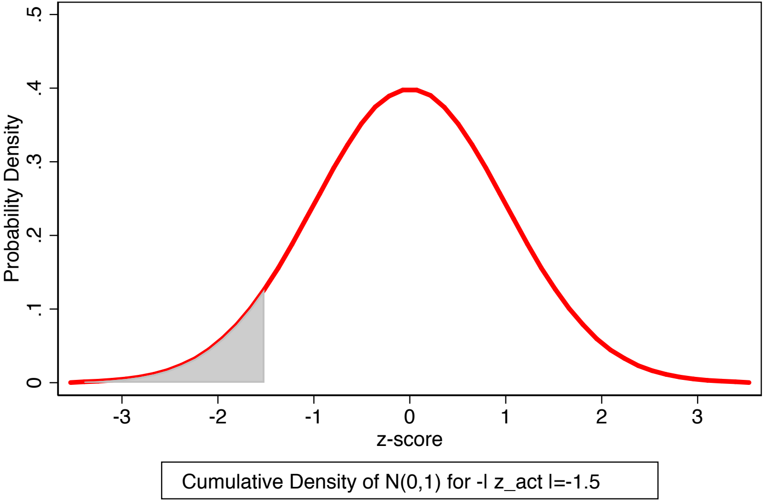
- The p-value for $\bar{Y}^{act}$ given a null ($H_0 : \mu_Y = \mu_{Y,0}$) and alternative ($H_1 : \mu_Y \neq \mu_{Y,0}$) hypothesis is defined as:

$$\text{p-value} = P(|z| > |z^{act}|) = 2 \times \Phi(-|z^{act}|)$$

  - Given $\mu_{Y,0}$, you can get a $z$ as large in magnitude as $|z^{act}|$ if $z$ is either really small or really large.
- The second equality comes from the fact that: (1) $z$ is distributed N(0,1) by the CLT if $n$ is large enough; and (2) the normal distribution is symmetric

# p-values Graphically

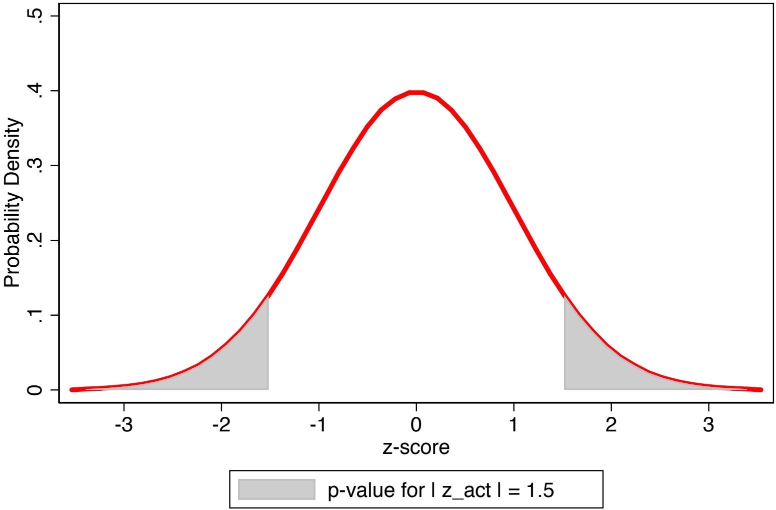
$|z^{act}| = 1.5$



Note: the N(0,1) cumulative density at -l z_act l is
the total amount of shaded grey area

# p-values Graphically

$|z^{act}| = 1.5$



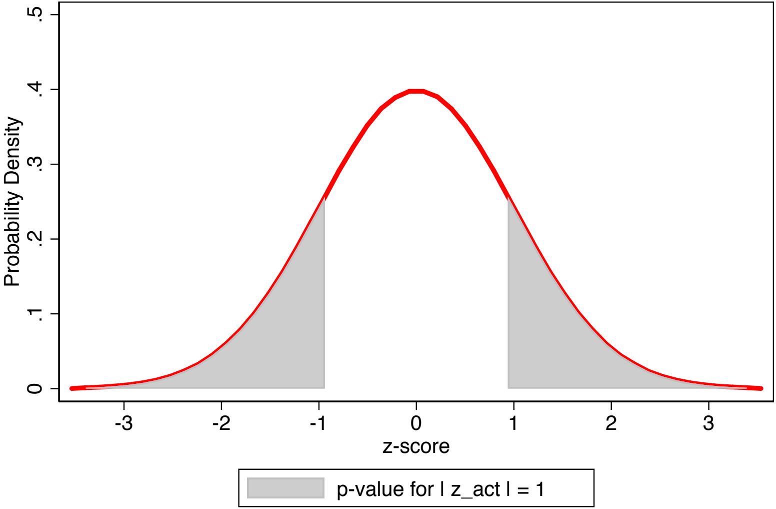
Note: the p-value for z_act is
the total amount of shaded grey area

# Interpreting p-values

- The p-value tells us the probability of obtaining a value as extreme as $\bar{Y}^{act}$ if the null $H_0 : \mu_Y = \mu_{Y,0}$ is true, given $\sigma_{\bar{Y}}$
- If a p-value for $\bar{Y}^{act}$ is small, then this means it was unlikely that we would have obtained as extreme a value of $\bar{Y}^{act}$ (either high or low) from a random sample from the population if the true value of the mean was indeed $\mu_Y = \mu_{Y,0}$
- Small p-values arise either because $\bar{Y}^{act} - \mu_{Y,0}$ is big or $\sigma_{\bar{Y}}$ is small (or both) implies we have little confidence from our sample in the null hypothesis that $H_0 : \mu_Y = \mu_{Y,0}$
- In contrast, the higher the p-value, the more likely (or there is a higher probability that) $\bar{Y}^{act}$ was generated by a population density with mean $\mu_{Y,0}$

# p-values Graphically

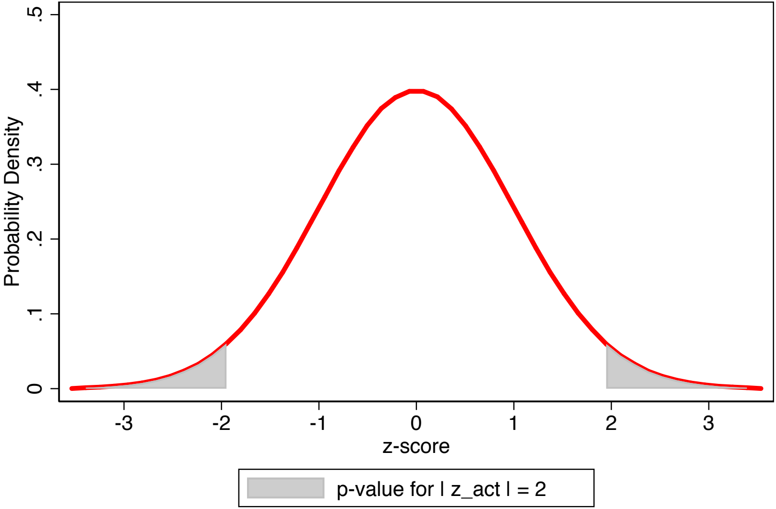
$\left(\bar{Y}^{act} - \mu_{Y,0}\right) \downarrow$ and $z^{act} \downarrow$: null hypothesis more likely to be true



Note: the p-value for z_act is
the total amount of shaded grey area

# p-values Graphically

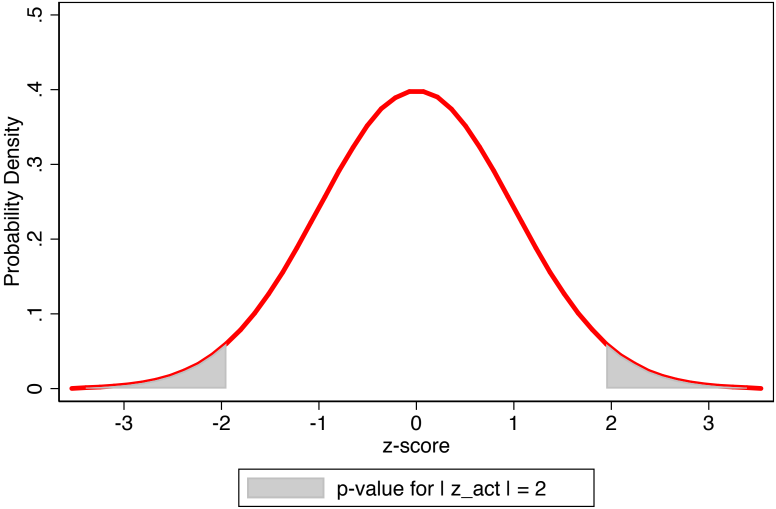
Either $\left(\bar{Y}^{act} - \mu_{Y,0}\right) \uparrow$ and $z^{act} \uparrow$: null hypothesis less likely to be true



Note: the p-value for z_act is
the total amount of shaded grey area

# p-values Graphically

Or $\sigma_{\bar{Y}} \downarrow$ and $z^{act} \uparrow$: null hypothesis less likely to be true



Note: the p-value for z_act is
the total amount of shaded grey area

# p-value Calculations

- Why do we use absolute values in computing a p-value?
- We use absolute values to capture the two ways in which $H_0 : \mu_Y = \mu_{Y,0}$ is false in favour of $H_1 : \mu_Y \neq \mu_{Y,0}$:
  - $\bar{Y}^{act} < \mu_{0,Y}$ is much <u>smaller</u> than the hypothesized $\mu_{0,Y}$ value
  - $\bar{Y}^{act} < \mu_{0,Y}$ is much <u>smaller</u> than the hypothesized $\mu_{0,Y}$ value
- So computing a p-value as

$$\text{p-value} = P(|z| > |z^{act}|) = 2 \times \Phi(-|z|)$$

is a short-cut calculation that that exploits the CLT, symmetry of the N(0,1), and accounts for the fact that we reject $H_0$ in favour of $H_1$ if we obtain really big or really small $\bar{Y}$ values from our sample relative to a population value $\mu_{Y,0}$

# Sample Variance and Sample Standard Deviation

$$z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$$

- To compute p-values based on our random sample, we need to be able to estimate $\sigma_{\bar{Y}}$ in the denominator
    - This is the "we will come back to this later" comment from slide 18 above
- Since $\sigma_{\bar{Y}} = \frac{\sigma_Y^2}{\sqrt{n}}$, we need to estimate $\sigma_Y^2$, the population variance for $Y$
- An unbiased and consistent estimate of the population variance $\sigma_Y^2$ is the sample variance, which we denote $s_Y^2$:

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

# Standard Error of the Sample Mean

- We can therefore estimate $\sigma_Y$ using $s_Y$
- Putting it all together, this means we can estimate $\sigma_{\bar{Y}}$ with:

$$SE(\bar{Y}) = \frac{s_Y}{\sqrt{n}}$$

where $SE(\bar{Y})$ is called the standard error of $\bar{Y}$

# t-statistics

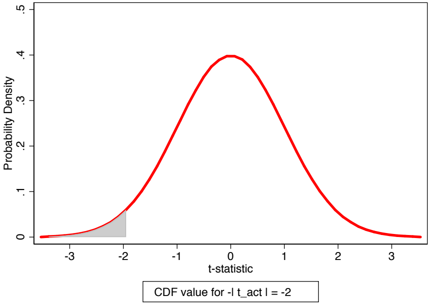
- The standardized sample average plays a central role in hypothesis testing and is called the t-statistic or t-ratio:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

- It is similar to the z-score except that the denominator is the (estimated) standard error of $\bar{Y}$, not the (known) $\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$
- Key result: if sample size $n$ is large, $t$ is distributed $N(0, 1)$
- This allows us to use t-statistics and the normal distribution together to conduct empirical hypothesis tests
- In practice, we use the sample average and standard error that we compute using our random sample, $\bar{Y}^{act}$ and $SE(\bar{Y}^{act})$, for computing t-statistics for a given hypothesized $\mu_{Y,0}$:

$$t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y})^{act}}$$

# Computing p-values with t-statistics

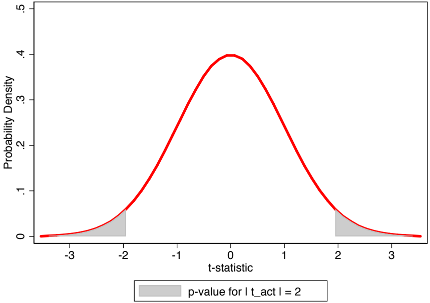


- $t$ has a N(0,1) distribution (if $n$ is large enough)
- Cumulative density function (CDF) of N(0,1) is denoted $\Phi(\cdot)$, which by the definition of the CDF implies

$$P(t \leq -|t^{act}|) = \Phi(-|t^{act}|)$$

- $P(t \leq -|t^{act}|)$ is shown in the grey area in the graph above

# Computing p-values with t-statistics



- Further recall the definition of a p-value:

$$\text{p-value} = P(|t| > |t^{act}|)$$

- As before, because N(0,1) is symmetric, we can use its CDF for testing the null $H_0 : \mu_Y = \mu_{Y,0}$ using $t$ statistics:

$$\text{p-value} = 2 \times \Phi(-|t^{act}|)$$

- The p-value for $t^{act}$ is the grey area in the graph above

# Hypothesis Testing with a Pre-specified Significance Level

- Suppose we want to have rule for rejecting (or not rejecting) our hypothesis $H_0 : \mu_Y = \mu_{Y,0}$ versus the alternative $H_1 : \mu_Y \neq \mu_{Y,0}$
- Example rejection rule: reject $H_0$ if the p-value is less than $\alpha = 5\%$, which leads to the rule:

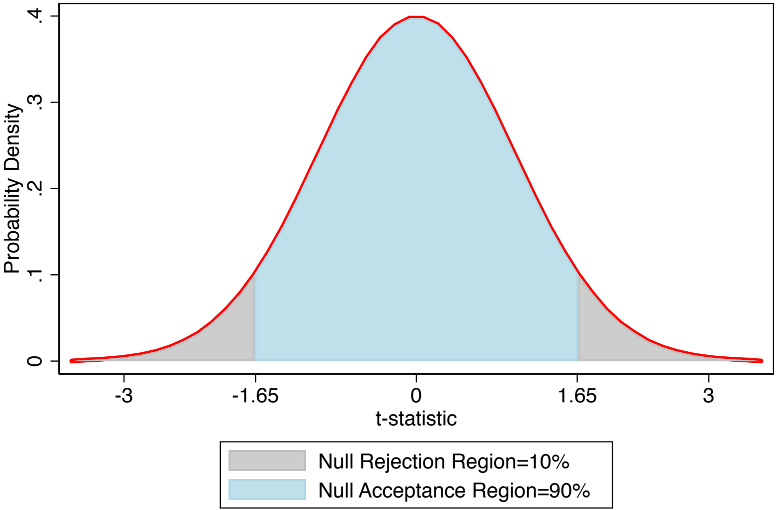
  $$\text{Reject } H_0 \text{ if } |t^{act}| > 1.96$$

  where 1.96 is the critical value for the test corresponding to $\alpha = 5\%$, which is often denoted by $t_\alpha$
- The value of 1.96 emerges because the area of the under the tails of the $N(0,1)$ distribution outside $\pm$ 1.96 is 5%

# Hypothesis Testing with a Pre-specified Significance Level

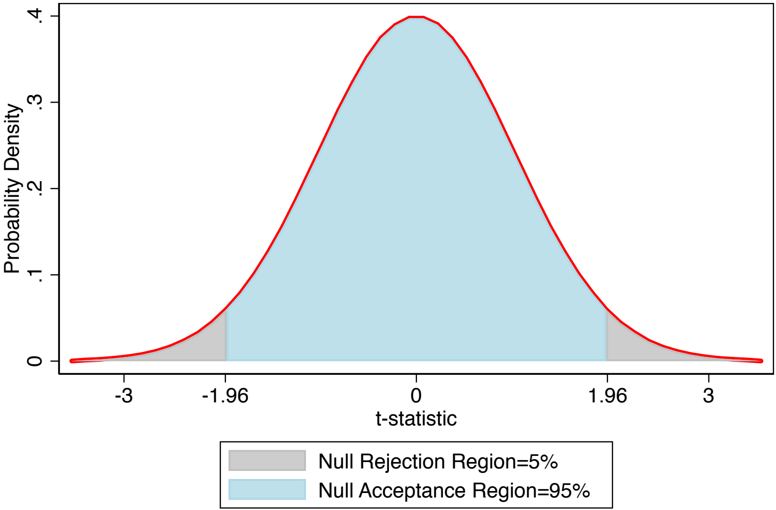
- $\alpha$ is called the level of significance
  - smaller $\alpha$ values implies a tougher test as the rejection region shrinks as $\alpha$ shrinks
  - typical values of $\alpha$ used are 10%, 5%, and 1% with corresponding critical values of 1.65, 1.96, and 2.58
- Interpretation: if $\alpha = 5\%$, the we reject the null $H_0 : \mu_Y = \mu_{Y,0}$ if the probability of obtaining a t-statistic as extreme $t^{act}$ is less than 5%
- Alternative interpretation: we are ok with the risk of incorrectly rejecting the null $H_0 : \mu_Y = \mu_{Y,0}$ when the null is in fact true 5% of the time given $t^{act}$
  - This is also called Type 1 Error

# Hypothesis Testing with $\alpha = 10\%$, $t_{0.10} = 1.65$



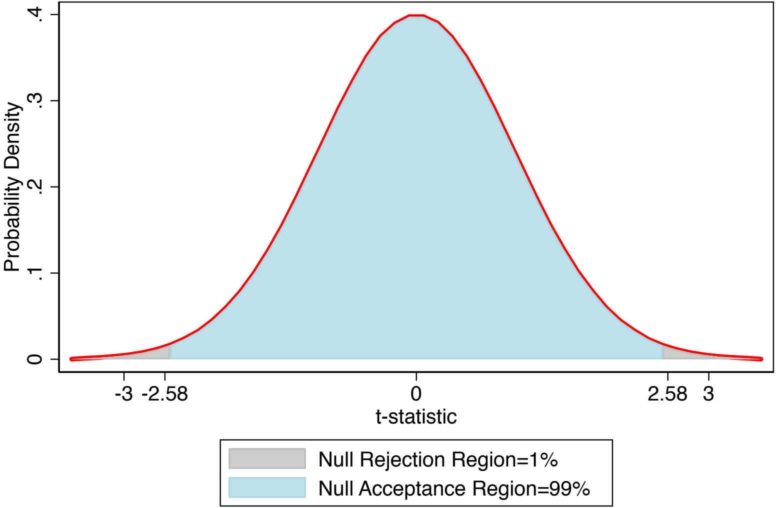
Rejection Rule: Reject Null if (t_act > 1.65 OR t_act < -1.65)

# Hypothesis Testing with $\alpha = 5\%$, $t_{0.05} = 1.96$



Rejection Rule: Reject Null if (t_act > 1.96 OR t_act < -1.96)

# Hypothesis Testing with $\alpha = 1\%$, $t_{0.01} = 2.58$



Rejection Rule: Reject Null if (t_act > 2.58 OR t_act < -2.58)

# Steps for Hypothesis Testing a Mean

1. State hypothesis test: $H_0 : \mu_Y = \mu_{Y,0}$, $H_1 : \mu_Y \neq \mu_{Y,0}$
2. Compute sample mean $\bar{Y}$:

$$\bar{Y} = \frac{\sum_i^n Y_i}{n}$$

3. Compute sample standard error of $\bar{Y}$:

$$SE(\bar{Y}) = \frac{s_Y}{\sqrt{n}}, \text{ where: } s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

4. Compute t-statistic:

$$t^{act} = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

# Steps for Hypothesis Testing a Mean

5. Compute p-value:

$$\text{p-value} = 2 \times \Phi(-|t^{act}|)$$

6. Specify a significance level $\alpha$ (say $\alpha = 5\%$) and use the rejection rule based on the critical value:

$$\text{Reject } H_0 \text{ if } |t^{act}| > 1.96 \text{ (for } \alpha = 5\%\text{)}$$

You can equivalently reject based on the p-value:

$$\text{Reject } H_0 \text{ if } 2\Phi(-|t^{act}|) < \alpha$$

- Simply put: big t-statistics or (equivalently) small p-values cause you to reject the null hypothesis

# Application of Hypothesis Testing to Pokies Data

- Politician says "we don't have a gambling problem in Australia – there's only 1 pokie per 1000 people on average!'
- We can use the statistical program R to test this hypothesis
  1. Hypothesis: $H_0 : \mu_Y = 1$, $H_1 : \mu_Y \neq 1$
  2. Sample mean:
  $$\bar{Y} = 4.88$$
  3. Standard error of mean:
  $$SE(\bar{Y}) = 0.22$$
  4. t-statistic:
  $$t^{act} = \frac{4.88 - 1}{0.22} = 17.63$$
  5. p-value is $< 0.00001$!
  6. Rejection rule at $\alpha = 5\%$ significance level:
  $$|t^{act}| = 17.63 > 1.96$$
  which implies we reject the null that $\mu_Y = 1$ at the 5% level of significance!

# Application of Hypothesis Testing to Pokies Data

- What if $H_0 : \mu_Y = 4$?
  - $\bar{Y} = 4.88$, $SE(\bar{Y}) = 0.22$
  - $t^{act}$=4.04
  - p-value=0.0001
  - $|t^{act}|$=4.04 $> 1.96$, so reject null at $\alpha = 5\%$

- What if $H_0 : \mu_Y = 4.5$?
  - $\bar{Y} = 4.88$, $SE(\bar{Y}) = 0.22$
  - $t^{act}$=1.74
  - p-value=0.0819
  - $|t^{act}|$=1.74 $< 1.96$, so fail to reject null at $\alpha = 5\%$

- Notice in the latter example that:
  - $|t^{act}|$=1.74 $> 1.65$, so reject null at $\alpha = 10\%$
  - Highlights how judgement calls about $\alpha$ can influence whether you reject or fail to reject a null

# One-Sided Alternatives

- We can also propose one-sided alternatives such as:

$$H_1 : \mu > \mu_{Y,0}$$

- In this case, we reject the null $H_0 : \mu = \mu_{Y,0}$ in favour of the alternative $H_1$ <u>only for really large</u> values of $t^{act}$
  - This is differs from how we reject the null with a two-sided alternative $H_1 : \mu \neq \mu_{Y,0}$, which we do for really large or really small values of $t^{act}$

# One-Sided Alternatives

- Given this, the p-value for the one-sided alternative $H_1 : \mu > \mu_{Y,0}$ is the area of the standard normal distribution to the right of $t^{act}$
- So the p-value for the t-statistic corresponds to t-statistics to the left of (less than) $t^{act}$) in the N(0,1) distribution :

  $$\text{p-value} = 1 - \Phi(t^{act})$$

  where note we compute $t^{act}$) exactly as we did for two-sided hypothesis tests
- The rejection rules are what change for one-sided hypothesis tests:
  - $\alpha = 0.10$: reject $H_0$ if $t^{act} > 1.28$
  - $\alpha = 0.05$: reject $H_0$ if $t^{act} > 1.65$
  - $\alpha = 0.01$: reject $H_0$ if $t^{act} > 2.33$

# One-Sided Alternatives

- Everything is just the opposite for a one-sided alternative with the less than inequality

$$H_1 : \mu < \mu_{Y,0}$$

- In this case, we reject the null $H_0 : \mu = \mu_{Y,0}$ in favour of the alternative $H_1$ $\underline{\text{only for really small}}$ values of $t^{act}$
- p-value based on the N(0,1) distribution for the t-statistic is:

$$\text{p-value} = \Phi(t^{act})$$

- The rejection rules for different levels of significant are:
  - $\alpha = 0.10$: reject $H_0$ if $t^{act} < -1.28$
  - $\alpha = 0.05$: reject $H_0$ if $t^{act} < -1.65$
  - $\alpha = 0.01$: reject $H_0$ if $t^{act} < -2.33$

# Confidence Intervals

- $\bar{Y}$ will never tell us exactly what $\mu_Y$ is because of random sampling
- However, we can use our sample to determine a set of values that contains that likely contains the true value of $\mu_Y$ with some confidence level
- We call this set a confidence interval for $\mu_Y$
- We construct the confidence interval using the t-statistic:

$$t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$$

where $\mu_{Y,0}$ is our hypothesized value of the population mean

# Constructing a Confidence Interval

- First, let us specify a confidence level $\alpha = 5\%$, with its associated critical value of 1.96
- Notice that, given a sample mean, the values of hypothesised population means $\mu_{Y,0}$ for which we cannot reject a two-sided test at the 5% level of significance are such that:

$$-1.96 \leq \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})} \leq 1.96$$

- If we re-arrange each of these inequalities we obtain:

$$\bar{Y} - 1.96SE(\bar{Y}) \leq \mu_{Y,0} \leq \bar{Y} + 1.96SE(\bar{Y})$$

# Constructing a Confidence Interval

$$\bar{Y} - 1.96SE(\bar{Y}) \leq \mu_{Y,0} \leq \bar{Y} + 1.96SE(\bar{Y})$$

- This is the set of hypothesised values of $\mu_{Y,0}$ for which we cannot reject that the population mean ($\mu_Y$) is the same as the sample average ($\bar{Y}$) at the 5% level of significance
- We call this the 95% confidence interval (or 95% CI) for $\mu_Y$.
- The 95% CI for $\mu_Y$ is often written like:

$$[\bar{Y} - 1.96SE(\bar{Y}), \bar{Y} + 1.96SE(\bar{Y})]$$

# Interpreting a Confidence Interval

- In a 95% CI for $\mu_Y$:

  $$[\bar{Y} - 1.96SE(\bar{Y}), \bar{Y} + 1.96SE(\bar{Y})]$$

  $\bar{Y} - 1.96SE(\bar{Y})$ is the lower bound of the interval, and $\bar{Y} + 1.96SE(\bar{Y})$ is the upper bound
- Interpretation: with random sampling, given values $\bar{Y}$ and $SE(\bar{Y})$, there is a 95% chance that the true value of the mean $\mu_Y$ sits between the lower and upper bound of the 95% CI
  - we fail to reject the null for hypothesised $\mu_Y$ values as small as $\bar{Y} - 1.96SE(\bar{Y})$
  - we fail to reject the null for hypothesised $\mu_Y$ values as large as $\bar{Y} + 1.96SE(\bar{Y})$

# Common Confidence Intervals

- 90% CI at the $\alpha = 10\%$ level of significance:

$$[\bar{Y} - 1.65 SE(\bar{Y}), \bar{Y} + 1.65 SE(\bar{Y})]$$

- 95% CI at the $\alpha = 5\%$ level of significance:

$$[\bar{Y} - 1.96 SE(\bar{Y}), \bar{Y} + 1.96 SE(\bar{Y})]$$

- 99% CI at the $\alpha = 1\%$ level of significance:

$$[\bar{Y} - 2.58 SE(\bar{Y}), \bar{Y} + 2.58 SE(\bar{Y})]$$

- General formula for (1-$\alpha$)% CI at $\alpha$ level of significance:

$$[\bar{Y} - t_\alpha SE(\bar{Y}), \bar{Y} + t_\alpha SE(\bar{Y})]$$

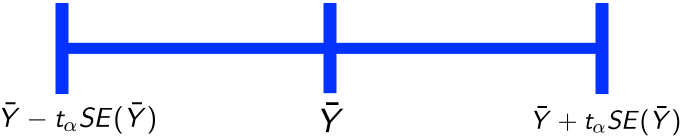
- CI's have 3 ingredients: $\bar{Y}$, $SE(\bar{Y})$, and the critical value $t_\alpha$ associated with specified level of significance $\alpha$

# Visualizing Confidence Intervals

## Base Case

- (1-$\alpha$)% CI:

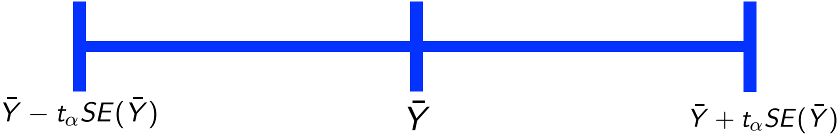
$$[\bar{Y} - t_\alpha SE(\bar{Y}), \bar{Y} + t_\alpha SE(\bar{Y})]$$


$\bar{Y} - t_\alpha SE(\bar{Y})$
$\bar{Y}$
$\bar{Y} + t_\alpha SE(\bar{Y})$

# Visualizing Confidence Intervals

## Impact of ↑ $(1-\alpha)$

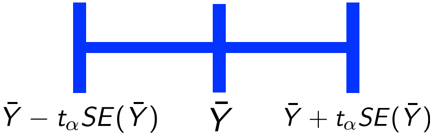
- (1-$\alpha$)% CI:
$$[\bar{Y} - t_\alpha SE(\bar{Y}), \bar{Y} + t_\alpha SE(\bar{Y})]$$
- Holding $SE(\bar{Y})$ fixed, ↑ $(1-\alpha)$ implies $t_\alpha$ ↑: we have a wider interval which increases our confidence that $\mu_{Y,0}$ lies within it
- For example, a 99% CI is wider than a 95% CI


$\bar{Y} - t_\alpha SE(\bar{Y})$
$\bar{Y}$
$\bar{Y} + t_\alpha SE(\bar{Y})$

# Visualizing Confidence Intervals

## Impact of ↓ $(1-\alpha)$

- (1-$\alpha$)% CI:
$$[\bar{Y} - t_\alpha SE(\bar{Y}), \bar{Y} + t_\alpha SE(\bar{Y})]$$
- Holding $SE(\bar{Y})$ fixed, $\downarrow (1-\alpha)$ implies $t_\alpha \downarrow$: we have a more narrow interval which decreases our confidence that $\mu_{Y,0}$ lies within it
- For example, a 90% CI is more narrow than a 95% CI

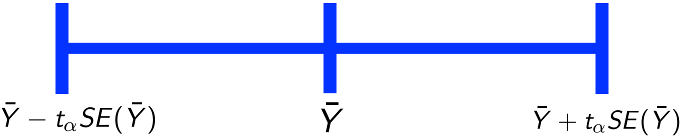

$\bar{Y} - t_\alpha SE(\bar{Y})$
$\bar{Y}$
$\bar{Y} + t_\alpha SE(\bar{Y})$

# Visualizing Confidence Intervals

## Back to the Base Case

- (1-$\alpha$)% CI:

$$[\bar{Y} - t_\alpha SE(\bar{Y}), \bar{Y} + t_\alpha SE(\bar{Y})]$$


$\bar{Y} - t_\alpha SE(\bar{Y})$
$\bar{Y}$
$\bar{Y} + t_\alpha SE(\bar{Y})$

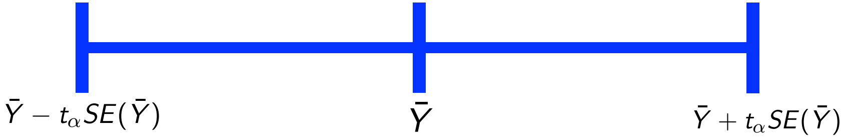
# Visualizing Confidence Intervals

## Impact of ↑ $SE(\bar{Y})$

- (1-$\alpha$)% CI:

$$[\bar{Y} - t_{\alpha} SE(\bar{Y}), \bar{Y} + t_{\alpha} SE(\bar{Y})]$$

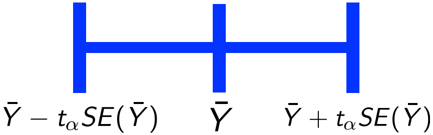
- Holding $\alpha$ fixed, ↑ $SE(\bar{Y})$: for a given $(1 - \alpha)$, a noiser estimate of the sample mean implies that we need a wider interval to achieve a $(1 - \alpha)$ confidence interval to ensure that $\mu_{Y,0}$ lies within the interval
- Less precision (↑ $SE(\bar{Y})$), say from noisier data or a smaller sample size $N$, forces us to be less confident that we can find $\mu_{Y,0}$ within narrower intervals


$\bar{Y} - t_{\alpha} SE(\bar{Y})$
$\bar{Y}$
$\bar{Y} + t_{\alpha} SE(\bar{Y})$

# Visualizing Confidence Intervals

## Impact of ↓ $SE(\bar{Y})$

- (1-$\alpha$)% CI:
$$[\bar{Y} - t_\alpha SE(\bar{Y}), \bar{Y} + t_\alpha SE(\bar{Y})]$$
- Holding $\alpha$ fixed, ↓ $SE(\bar{Y})$: for a given $(1-\alpha)$, a more precise estimate of the sample mean implies that can achieve a $(1-\alpha)$ level of confidence that $\mu_{Y,0}$ lies within and interval using a more narrow interval
- More precision (↓ $SE(\bar{Y})$), say from less noisy data or a larger sample size $N$, allows us to be more confident that we can find $\mu_{Y,0}$ within narrower intervals


$\bar{Y} - t_\alpha SE(\bar{Y})$
$\bar{Y}$
$\bar{Y} + t_\alpha SE(\bar{Y})$

# Confidence Intervals Example

- From our pokies example, 95% CI for $\mu_Y$ is

  $$[4.87 - 1.96 \times 0.22, 4.87 + 1.96 \times 0.22]$$

  which is

  $$[4.44, 5.30]$$

- Interpretation: There is a 95% chance that the true value of the mean number of pokies per 1,000 people is between 4.44 and 5.30
- The 90, 95, and 99% CIs are:
  - 90% CI $=$ [4.51, 5.23]
  - 95% CI $=$ [4.44, 5.30]
  - 99% CI $=$ [4.30, 5.44]
- Notice how the CI gets wider if $\alpha \downarrow$
  $\rightarrow$ you need a wider interval to be more confident that that $\mu_Y$ is within it

# Comparing Means from Different Populations

- How do we conduct hypothesis tests to address questions like "do males and females make the same earnings on average"?
- Keeping with the example, to test the difference between two means, let $\mu_w$ be hourly earnings for female college graduates (women), and $\mu_m$ be hourly earnings for male college graduates
- Let the hypothesised value of the difference be $d_0$
- The null and two-sided alternative is:

$$H_0 : \mu_w - \mu_m = d_0; \quad vs. \quad H_1 : \mu_w - \mu_m \neq d_0$$

- A common value for $d_0$ is $d_0 = 0$
  - In words: we test the null that males and females have the same earnings vs the alternative that they do not

# Comparing Means from Different Populations

- As before, we use the sample averages to estimate $\mu_w$ and $\mu_m$, $\bar{Y}_w$ and $\bar{Y}_m$
- Under the CLT, $\bar{Y}_w$ is distributed $N(\mu_w, \sigma_w^2/n_w)$ where $\sigma_w^2$ is the population variance of female earnings, and $n_w$ is the number of females in the sample
- Similarly, $\bar{Y}_m$ is distributed $N(\mu_m, \sigma_m^2/n_m)$ where $\sigma_m^2$ is the population variance of male earnings, and $n_m$ is the number of males in the sample
- Because $\bar{Y}_w$ and $\bar{Y}_m$ are constructed from independent random samples, the distribution of $\bar{Y}_w - \bar{Y}_m$ is

$$N(\mu_w - \mu_m, (\sigma_w^2/n_w) + (\sigma_m^2/n_m))$$

which comes directly from our rules for expectations and variances of sums of random variables, and independence

# Comparing Means from Different Populations

- Distribution of $\bar{Y}_w - \bar{Y}_m$ is

$$N(\mu_w - \mu_m, (\sigma_w^2/n_w) + (\sigma_m^2/n_m))$$

- As before, we can consistently estimate $\sigma_w^2$ and $\sigma_w^2$ with the sample variance $s_w^2$ and $s_w^2$
- So the standard error of $\bar{Y}_w - \bar{Y}_m$ is:

$$SE(\bar{Y}_w - \bar{Y}_m) = \sqrt{s_w^2/n_w + s_m^2/n_m}$$

- To test our hypothesis, we can compute the t-statistic exactly as before:

$$t^{act} = \frac{(\bar{Y}_w - \bar{Y}_m) - d_0}{SE(\bar{Y}_w - \bar{Y}_m)}$$

  which is distributed N(0,1) if $n_w$ and $n_m$ are large enough

# Comparing Means from Different Populations

- p-value for $t^{act}$ is $2 \times \Phi(-|t^{act}|)$
- We reject the null at the $\alpha = 0.10, 0.05, 0.01$ levels of significance if $|t^{act}| > 1.65$ or $|t^{act}| > 1.96$ or $|t^{act}| > 2.58$
- Testing one-sided alternatives also follow exactly as before
- Confidence interval for the difference in means is computed using the t-statistic from the two-sided hypothesis test
- The 95% confidence interval for $d$ consists of those values of $d$ that are within $\pm$ 1.96 standard errors of $\bar{Y}_w - \bar{Y}_m$:

  $$[(\bar{Y}_w - \bar{Y}_m) - 1.96 \times SE(\bar{Y}_w - \bar{Y}_m), (\bar{Y}_w - \bar{Y}_m) + 1.96 \times SE(\bar{Y}_w - \bar{Y}_m)]$$

# Comparing Means with Randomised Control Trials

- Economists are increasingly using randomised control trials (RCTs) to test theories and policies
- The idea mimics what drug companies do
  - find a sample of sick patients
  - randomly give some patients a drug
  - to determine if the drug works, compare health outcomes from those who were given the drug to those who were not
- Let $\bar{Y}_T$ (treated) be the sample mean of health for people who were given the drug, and $\bar{Y}_{NT}$ (not treated) be the sample mean of health for people who were not given the drug
- We can use tests of means from different populations just discussed to test the hypothesis that $\mu_T - \mu_{NT} = 0$; that is, the null that the drug has no causal effect on health
- We will causal effects and experiments around week 9 and 10

# Using the $t$-statistic When Sample Size is Small

- Throughout, we have been assuming large sample sizes $n$ such that the distribution of $t$-statistics is well-approximated by a N(0,1) distribution
- What do we do if $n$ is small and the N(0,1) approximation is potentially poor?
- If the underlying distribution of $Y$ is normally distributed, then $t^{act} = \frac{\bar{Y}^{act} - \mu_{Y,0}}{SE(\bar{Y}^{act})}$ has a Student-t distribution, with $n-1$ degrees of freedom
- The Student-t distribution with $n-1$ degrees of freedom can be used instead of the N(0,1) distribution to compute p-values and construct confidence intervals if $n$ is small

# Using the $t$-Statistic When Sample Size is Small

- Modifications are also needed for t-statistics for testing differences of means from two random samples if $n$ is small; see the text on p. 134 and 135 for details if interested
- However, in practice we rarely work with samples in business or economics where $n$ is so small that we need to worry about $N(0,1)$ not being a good approximation for the distribution of t-statistics
- Going forward we only focus on doing hypothesis tests assuming sufficiently large sample size that allows us to use N(0,1) approximations for distributions of t-statistics

Note: skipping t-statistic for testing differences of means with small smaples on p.134-135. See the text for details. Again, not often used in practice.

# Scatterplots, Sample Covariance, Sample Correlation

- The final statistical calculations of note are the sample covariance and the sample correlation
- Sample covariance between random variables $X$ and $Y$ :

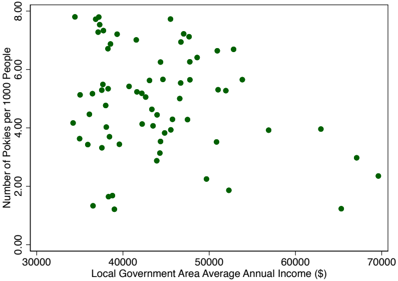
$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

- Sample correlation (or correlation coefficient):

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

where $-1 \leq r_{XY} \leq 1$

# Pokies Per 1000 People vs. LGA Average Income



- Sample Covariance is -3.22; Sample Correlation is -0.237
- Questions from the Prime Minister:
  - Is this relationship real? That is, could it just occur because of random sampling/chance?
  - And is this relationship concerning? That is, does an LGA having lower income cause it to have more pokies per person?
- To address these questions and other important economic questions, we need to understand regression analysis

# Summary of Population Values and Sample Counterparts

| Name | Population<br>Value | Sample<br>Estimator | Estimator<br>Calculation |
|---|---|---|---|
| Mean | $\mu_Y$ | $\bar{Y}_Y$ | $\frac{\sum_{i=1}^{N} Y_i}{N}$ |
| Variance | $\sigma^2_X$ | $s^2_Y$ | $\frac{1}{n-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$ |
| Standard Deviation | $\sigma_Y$ | $s_Y$ | $\sqrt{s^2_Y}$ |
| Standard Error of $\bar{Y}$ | $\sigma_{\bar{Y}}$ | $SE(\bar{Y})$ | $\frac{s_Y}{\sqrt{n}}$ |
| Covariance | $\sigma_{XY}$ | $s_{XY}$ | $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ |
| Correlation | $\rho_{XY}$ | $r_{XY}$ | $\frac{s_{XY}}{s_X s_Y}$ |

# Summary of Testing

- Test statistics
  - z-score: $z = \frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$ (assumes $\sigma_{\bar{Y}}$ is known)
  - t-statistic: $t = \frac{\bar{Y} - \mu_{Y,0}}{SE(\bar{Y})}$ (this is what we actually use)
- Two-sided hypothesis test $(=)$ (this is what we most often do)
  - $H_0 : \mu_Y = \mu_{Y,0}$ versus alternative $H_1 : \mu_Y \neq \mu_{Y,0}$
  - p-value$=P(|t| > |t^{act}|) = 2 \times \Phi(-|t^{act}|)$
- One-sided hypothesis test $(>)$
  - $H_0 : \mu_Y = \mu_{Y,0}$ versus alternative $H_1 : \mu_Y > \mu_{Y,0}$
  - p-value$=1 - \Phi(|t^{act}|)$
- One-sided hypothesis test $(<)$
  - $H_0 : \mu_Y = \mu_{Y,0}$ versus alternative $H_1 : \mu_Y < \mu_{Y,0}$
  - p-value$=\Phi(|t^{act}|)$
- For all hypothesis tests (e.g., $=, >, <$): reject $H_0$ if p-value $<$ $\alpha$ where $\alpha$ is the user-specified significance level
  - can equivalently reject $H_0$ if $|t^{act}| > t_\alpha$
  - commonly-used $\alpha$ values are 0.01, 0.05, 0.10

# Summary of Confidence Intervals

- $(1-\alpha)$=99% CI ($\alpha = 0.01$ significance level):

$$[\bar{Y} - 2.58SE(\bar{Y}), \bar{Y} + 2.58SE(\bar{Y})]$$

- $(1-\alpha)$=95% CI ($\alpha = 0.05$ significance level):

$$[\bar{Y} - 1.96SE(\bar{Y}), \bar{Y} + 1.96SE(\bar{Y})]$$

- $(1-\alpha)$=90% CI ($\alpha = 0.10$ significance level):

$$[\bar{Y} - 1.65SE(\bar{Y}), \bar{Y} + 1.65SE(\bar{Y})]$$

- Confidence intervals are centered and symmetric around $\bar{Y}$ and become wider if $(1-\alpha)\uparrow$ or $SE(\bar{Y})\uparrow$