SOC40830 Quantitative Data Analytics and Applications

Dr. Eoin Flaherty (D413, Newman Building)
eoin.flaherty@ucd.ie

Week 5, Wednesday October 12th

1

Week 5 Outline

1. **Interpreting summary statistics: mean, median, mode, standard deviation, range, quantile.**
2. **The logic of statistical inference and models.**
3. **Independent group comparison: the t-test.**
4. **Tabular inference: the chi-square.**

2

1. Interpreting summary statistics

Mean, Median, and Mode

Central tendency/typical value, and *spread/dispersion*

Arithmetic Mean: sum of all cases divided by number of cases.

Median: middle value of a rank-ordered distribution.

Mode: most frequently occurring value in a distribution.

3

1. Interpreting summary statistics

Median and Range

For a given variable, the **median** is the middle value of a **rank-ordered** set of observations.

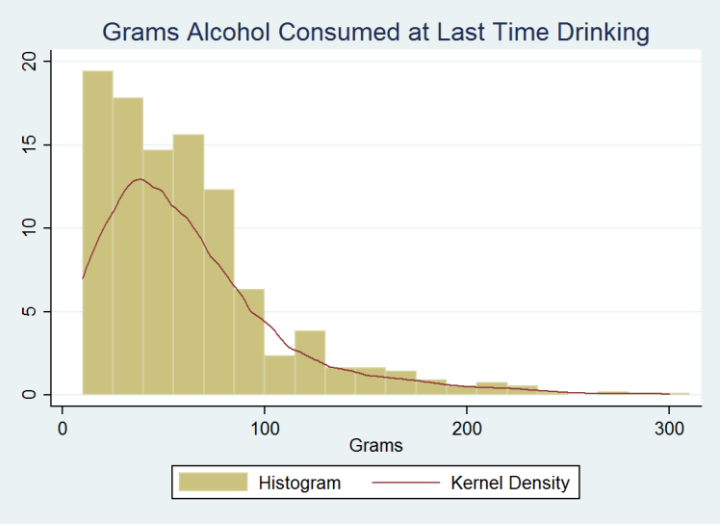
It is also denoted as the 0.5 quantile, or 50^{th} percentile.

The **range** is the difference between the highest and lowest case.

4

1. Interpreting summary statistics

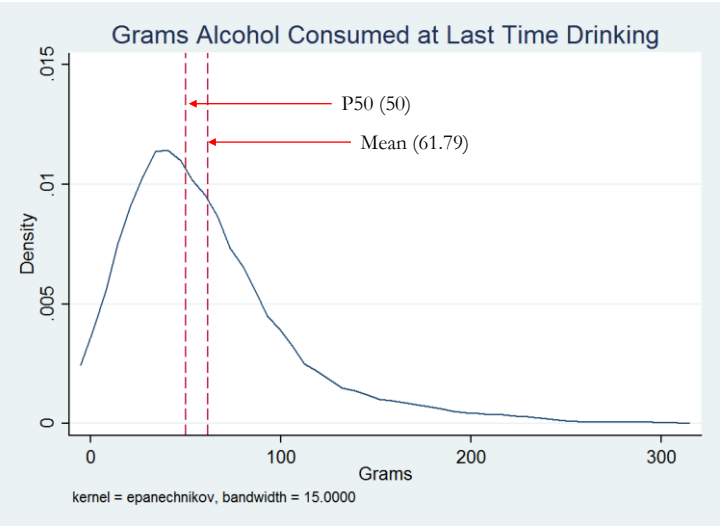
```
twoway (hist alcwkdy if alcwkdy<=300, percent width(15) xtitle(Grams))/*
*/(kdensity alcwkdy if alcwkdy<=300, bwidth(15) area(1124)/*
*/title(Grams Alcohol Consumed at Last Time Drinking))
```



5

1. Interpreting summary statistics

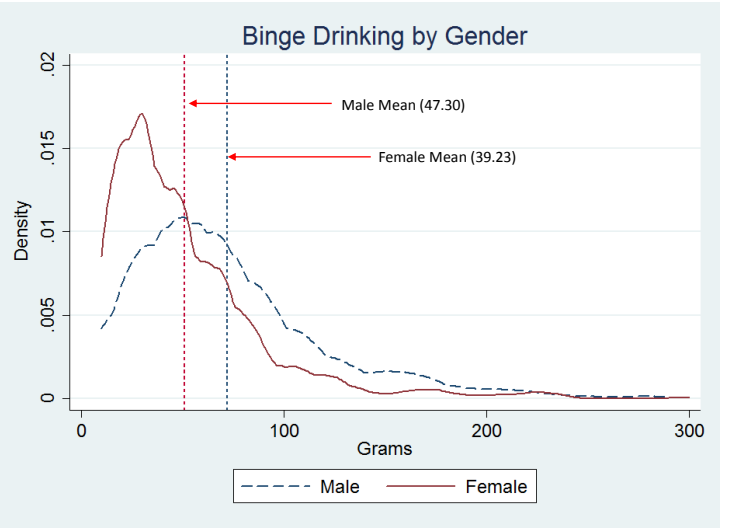
```
tabstat alcwkdy if alcwkdy<=300, s(min p50 max count mean sd)
kdensity alcwkdy if alcwkdy<=300, bwidth(15) xline(50 61.79) /*
*/title(Grams Alcohol Consumed at Last Time Drinking)
```



6

1. Interpreting summary statistics

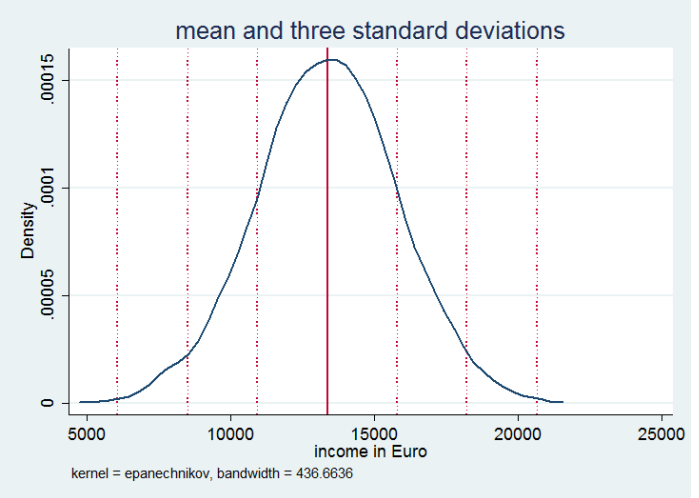
```
twoway (kdensity alcwkdy if alcwkdy<=300 & gndr==1, xline(72.18) /*
*/title(Binge Drinking by Gender))/*
*/ (kdensity alcwkdy if alcwkdy<=300 & gndr==2, xline(51.26))
```



7

1. Interpreting summary statistics

```
kdensity x2, xline(`truemean') xline(`Xplus3SE' `Xminus3SE'  `XplusSE' /*
*/`XminusSE' `Xplus2SE' `Xminus2SE', lpattern(dot))/*
*/ title(mean and three standard deviations) xtitle(income in Euro)
```



Each band denotes one standard deviation.

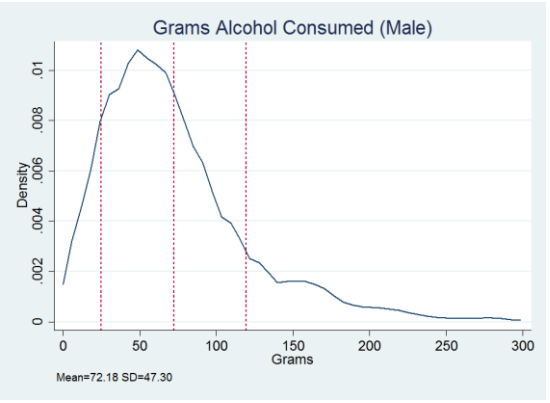
- Approximately 68% of cases fall within one standard deviation of the mean
- Approximately 95% of cases fall within one standard deviation of the mean
- Approximately 99.7% of cases fall within one standard deviation of the mean

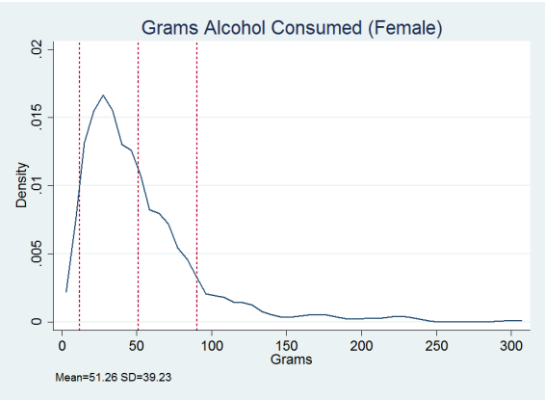
8

1. Interpreting summary statistics

```
tabstat alcwkdy if alcwkdy<=300 & gndr==1, /*
*/s(min p50 max count mean sd)
kdensity alcwkdy if alcwkdy<=300 & gndr==1, /*
*/xline(24.88 72.18 119.48) xlabel(0(50)300) xtitle(Grams)/*
*/ title(Grams Alcohol Consumed (Male)) /*
*/note(Mean=72.18 SD=47.30)
```

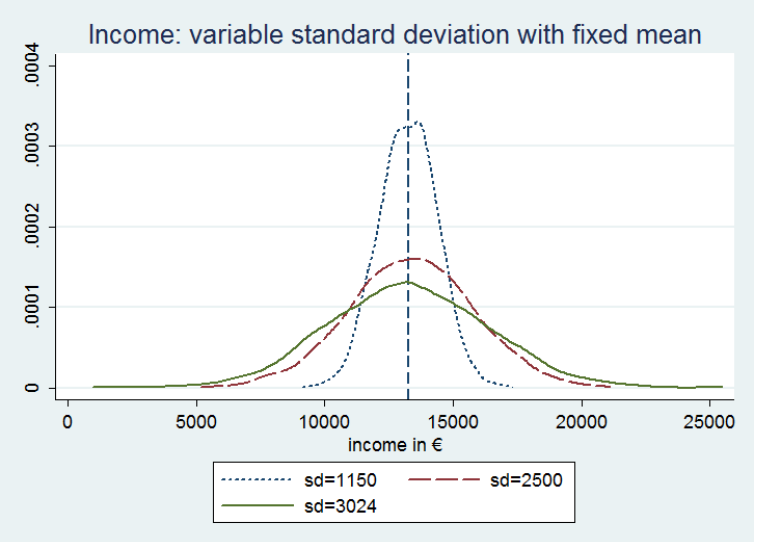
```
tabstat alcwkdy if alcwkdy<=300 & gndr==2, /*
*/s(min p50 max count mean sd)
kdensity alcwkdy if alcwkdy<=300 & gndr==2, /*
*/xline(12.03 51.26 90.49) xlabel(0(50)300) xtitle(Grams)/*
*/ title(Grams Alcohol Consumed (Female)) /*
*/note(Mean=51.26 SD=39.23)
```





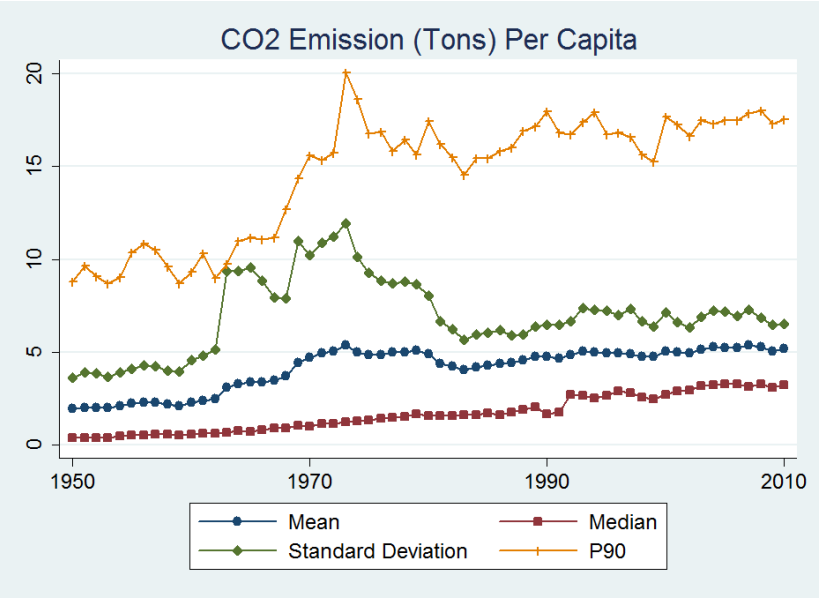
9

1. Interpreting summary statistics



10

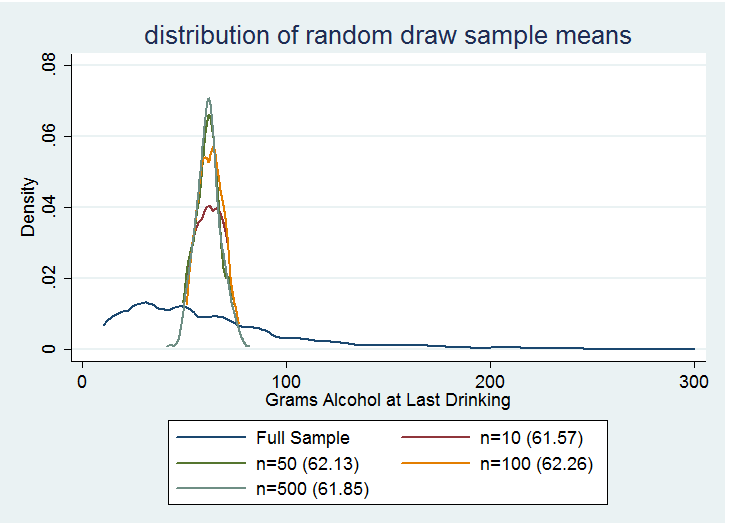
1. Interpreting summary statistics



11

2. The Logic of Statistical Inference

Sampling Distribution and Standard Error



12

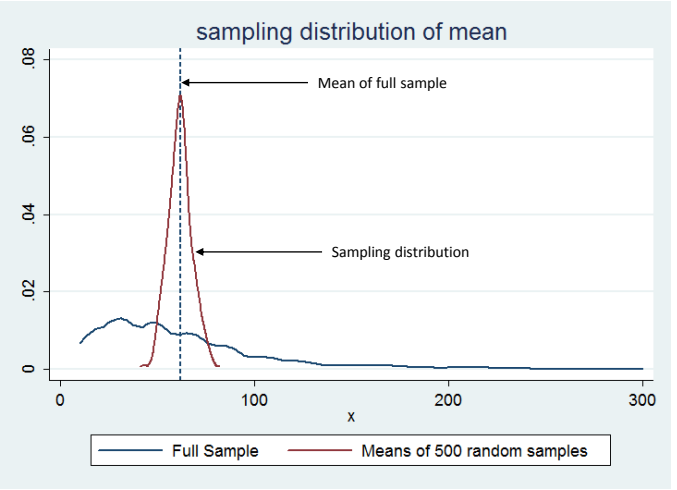
2. The Logic of Statistical Inference

Sampling Distribution and Standard Error

- Sampling distribution: distribution of a statistic in infinite number of random samples.
- The mean of sample means for an infinite number of random sample is equal to the population mean (central limit theorem – attractor distributions). A large set of samples also has a normal distribution.
- Key question: is our sample statistic (point estimate) close to the 'true' (population) value, or far off?
- The more the point estimate varies around the true mean, the less confidence we have in the accuracy of our sample statistic.
- The sampling distribution is the basis for establishing our level of trust.

13

2. The Logic of Statistical Inference



The standard error

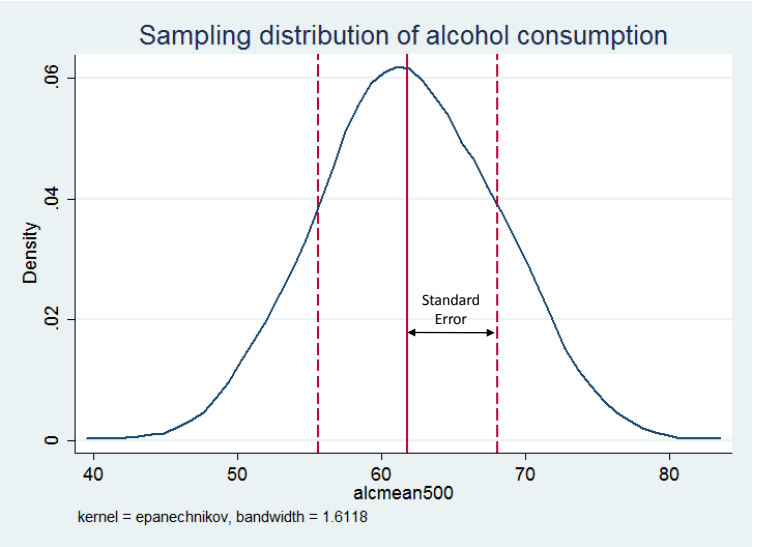
The standard error is the standard deviation of the sampling distribution (6.21). In the example on the left, the distribution of 500 sample means is plotted.

This gives an indication of the variability of a point estimate (the mean) across a succession of random draws.

Mean alcohol consumption in repeat random samples therefore deviates 6.21 grams on average from 'true' value.

14

2. The Logic of Statistical Inference



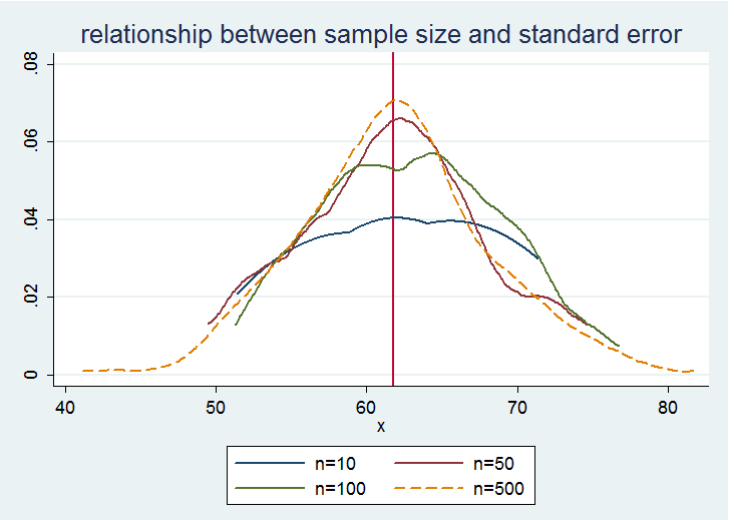
The standard error

- Means closer to true value more frequent.
- Approx 69% of sample means within one SD of mean.
- Approx 95% of sample means within two SD of mean.

```
sum alcwkdy if alcwkdy<=300
local truemean=r(mean)
sum alcmean500
local XplusSE = r(mean) + r(sd)
local XminusSE = r(mean) - r(sd)
kdensity alcmean500,
xline(`truemean') /*
*/xline(`XplusSE' `XminusSE',
lpattern(dash))
```

15

2. The Logic of Statistical Inference



The standard error

- Larger samples produce narrower sampling distributions, and smaller standard errors.
- For larger samples, the point estimates group closer to the true mean.

```
sum alcwkdy if alcwkdy<=300
local truemean=r(mean)
twoway kdensity alcmean10 if alcwkdy<=300/*
*/|| kdensity alcmean50 if alcwkdy<=300/*
*/|| kdensity alcmean100 if alcwkdy<=300/*
*/|| kdensity alcmean500 if alcwkdy<=300,
xline(`truemean') lpattern(dash)
```

16

2. The Logic of Statistical Inference

Sampling Distribution and Standard Error

- In practice, we use the sample data to estimate the standard error: $SE = \frac{SD(X)}{\sqrt{n}}$
- These calculations often need adjustment for features of the sampling technique.

```
. mean alcwkdy lrscale agea weight

Mean estimation                     Number of obs    =       900

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     alcwkdy |   63.76556   1.809454      60.21431     67.3168
     lrscale |   5.028889   .0638954      4.903487     5.15429
        agea |   49.69889    .572912      48.57449    50.82329
      weight |   74.91911   .5072541      73.92357    75.91465
--------------------------------------------------------------
```

17

2. The Logic of Statistical Inference

Sampling Distribution and Standard Error

- In practice, we use the sample data to estimate the standard error: $SE = \frac{SD(X)}{\sqrt{n}}$
- These calculations often need adjustment for features of the sampling technique and nonresponse (hint: we can deal with this through extensions to regression commands).

```
. mean alcwkdy if alcwkdy<=300

Mean estimation                     Number of obs    =      1124

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     alcwkdy |    61.7927    1.33339      59.17649    64.40892
--------------------------------------------------------------
```

18

2. The Logic of Statistical Inference

Confidence intervals as an indicator of point estimate reliability

95% of cases fall within 1.96 deviations of the mean: $\bar{y} \pm z(SE)$

$$61.79 + \big(1.96(1.33)\big) = 64.39$$
$$61.79 - \big(1.96(1.33)\big) = 59.18$$

Interpretation: 95% of intervals formed from this method contain population mean

```
Mean estimation                     Number of obs    =      1124

------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
     alcwkdy |    61.7927    1.33339      59.17649    64.40892
------------------------------------------------------------------
```

19

2. The Logic of Statistical Inference

Confidence intervals as an indicator of point estimate reliability

95% of cases fall within 1.96 deviations of the mean: $\bar{y} \pm z(SE)$

Interpretation: 95% of intervals formed from this method contain population mean

```
Mean estimation                     Number of obs    =      2171

------------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
        agea |   49.27637   .3877354        48.516    50.03674
      weight |   73.97771   .3256146      73.33916    74.61626
------------------------------------------------------------------
```

20

2. The Logic of Statistical Inference

Statistical Significance (Kohler and Kreuter p233)

- Significance test proposes some parameter in population=0, and calculates probability of observing value as large in our sample if that were true.
- Statistical significance ≠ substantive significance, or importance, or effect size – must be established by theory and contextual knowledge.
- We can formalise this into null and alternate hypotheses:

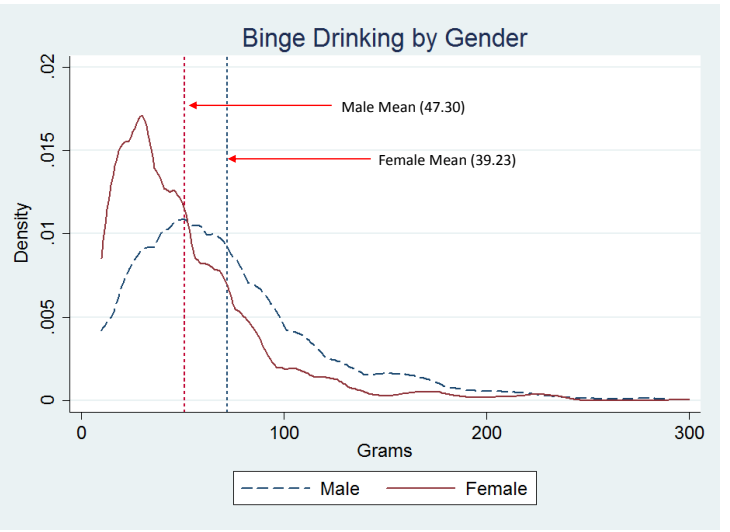
Ho: difference in alcohol consumption between males and females is 0

Ha: difference in alcohol consumption between males and females is non-0

21

3. Independent group comparison: the t-test

```
twoway (kdensity alcwkdy if alcwkdy<=300 & gndr==1, xline(72.18) /*
*/title(Binge Drinking by Gender))/*
*/ (kdensity alcwkdy if alcwkdy<=300 & gndr==2, xline(51.26))
```



22

3. Independent group comparison: the t-test

```
. mean alcwkdy if alcwkdy<=300, over(gndr)

Mean estimation                     Number of obs    =    1124

         Male: gndr = Male
       Female: gndr = Female

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
alcwkdy      |
        Male |   72.17668   1.988047      68.27597    76.07738
      Female |   51.25986   1.660859      48.00112    54.51859
--------------------------------------------------------------
```

23

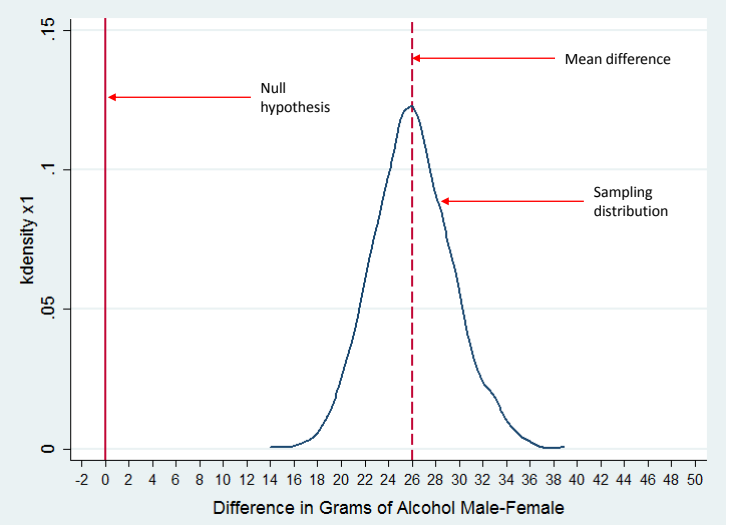
3. Independent group comparison: the t-test

Independent samples t-test / two-group mean comparison

- Takes the standard error of the difference between means.
- Asks how likely to observe a value of 26.03 if true difference is 0.
- The sampling distributions of differences between sample means are t-distributed: $t = \frac{\bar{X}_2 - \bar{X}_1}{se}$

24

3. Independent group comparison: the t-test



25

3. Independent group comparison: the t-test

```
. mvdecode gndr, mv(9=.a)
. ttest alcwkdy, by(gndr)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    Male |     577    78.35355    2.707761    65.04264    73.03526    83.67184
  Female |     560    52.32321    1.818926    43.04365    48.75045    55.89598
---------+--------------------------------------------------------------------
combined |    1137    65.53298    1.684497    56.80027    62.22791    68.83806
---------+--------------------------------------------------------------------
    diff |            26.03034    3.281108                19.59262    32.46806
------------------------------------------------------------------------------
    diff = mean(Male) - mean(Female)                              t =   7.9334
Ho: diff = 0                                     degrees of freedom =     1135

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

26