STAD 2012

1 a Categorical/nominal - no order of quantity involved - gender/day of week etc.

Quantive - quantity involved - underlying scale height/salary

b Graph of lenght plotted against frequency
- Bars all the same width
- Shows us the distribution of variable height

c Mean is the sum of the values of the variables divided by the amount of variables - it is a measure of centre or average

median is the value in the direct middle when the values are ordered - also a measure of centre

Expect them to be the same if the data is normally distributed implying that they are of similar value

d Standard deviation - Shows how much variation exists from the average (mean) or expected value low standard deviation indicates the data point tend to be very close to the mean
It is the positive square root of the variance

Percentile is the value of a variable below which a certain percent of observations fall eg. 20th percentile is the value below which 20 percent of observations can be in

e mean = 141 standard dev = 2

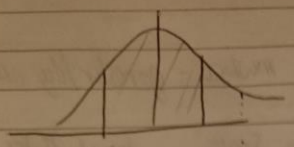
$141 - 138 = 3$ $-\frac{3}{2} < x < \frac{3}{2}$

$143 - 141 = 2$ $1.5 < x < 1$

$1 - 0.93 = 0.07$

$1 - 0.84 = 0.16$

=



$P(x < 1) - P(x < -1.5)$
0.84

$P(x$

$1 - P(x < 1.5)$

$1 - 0.93 = 0.7$

$0.84 - 0.7 = 0.77$ - probability

2a $1 \pm 1.96\sigma \frac{SD}{\sqrt{651}}$ $\frac{183}{651}$ purchased organic

$p = \frac{183}{651}$ $SE(p) = \sqrt{\frac{\frac{183}{651}\left(1-\frac{183}{651}\right)}{651}} = 0.017618826$

$\frac{183}{651} \pm 1.96(0.0176)$

± 0.0345329

$(0.25, 0.32)$ between 25% and 34% buy organic

b Card type 0 associated with buying organic.
Card type 0 not associated with buying organic

c Expected value - the weighted average of all possible values that the random variable can take on.

$(\pm)^{\%}$ % time Silver occurs x % time gold occurs.

$\frac{246}{651}$ $\frac{183}{651}$

time silver occurs multiplied x the gold silver occurs

d $DF = 3$ chisqur pdf = 11.834 between 5% and 2.5%.
Pvalue = 0.008 inside change of accuracy

do not which 4 suitable value of 0.05.
not reject null hypothesis do not be either 0.025 and 0.00 4

If p is less than confidence level we accept hypo
df = 198
$1.97 \leq t \leq 1.97$
t is outside of 1.97 so evidence against H_0
two long term averages are different

- $p = 0.0001$ which is $p < 0.05$
- Evidence against H_0

- 95% interval is 0.5045 and 0.3815 slope gets

3c. I would divide 347 by 16 get 21 I would then take every 21st number.

-Scatterplot

4a -Independent variable storage time on x-axis
- Dependent variable moisture content on y axis
- Predict the moisture content by storage time
- Used to find correlation between two variables

b. Equation to find the moisture at a certain time is
2.82 + 0.045 Storage time
- Coef 2.820 and estimate for first coefficient in model equation
- Coef 0.045 " "
- 2.820 moisture with no time storage
- 0.045 - Increase in Y per unit increase in x
- P value on storage time tests the hypothesis:
  H_0 Population Slope = 0
  H_1 Population Slope ≠ 0
  $p = < 0.001$ reject H_0
- P value on moisture tests:
  H_0 Population intercept = 0
  H_1 Population intercept ≠ 0
  $p = < 0.001$ reject H_0

STATS 2012

Q 3 a) The box was drawn by graphing one variable against another (graphing "no" against "right") the figure representing "no" and "right" were selected and then graphed

b) 95% confidence interval - 95% of the data

mu (no) - mu(yes)

$\bar{u}$ (no) - $\bar{u}$ (yes)

129.720 - 130.193 = -0.4730

We can conclude that between (-0.5647 and -0.3915 ~~the~~ ~~fraction~~ the mean of (no) is between (-0.5647, -0.3915) smaller than mean of (yes)

- The t-test assesses whether the means of two groups are different.
- measured by ratio: $\frac{\text{difference between two means}}{\text{measure of variability or dispersion of scores}}$

$$\rightarrow \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} \text{ (standard error)}$$

$$\rightarrow SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{Var_1}{n_1} + \frac{Var_2}{n_2}}$$ Var = variance, n = sample size

$$\rightarrow \text{variance} = (\text{standard deviation})^2$$

$$\rightarrow \text{Final formula } \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{Var_1}{n_1} + \frac{Var_2}{n_2}}}$$

$$\frac{-0.4730}{\sqrt{\frac{(10.3284)^2}{100} + \frac{(10.3284)^2}{100}}}$$

$$\frac{-0.473}{0.04633} = -10.14 \approx -10.20$$

- If t-value is positive → first mean larger, vice versa
- Check alpha level (0.05) and DF = $N_1 + N_2 - 2$
- look up t table row.

- P value < 0.001 ; < 0.5 Accept null hypothesis long term means are ~~not~~ same

5

STATS 2012

(b) - S = 0.013
- S is the standard deviation of the residual

- Residuals should be normally distributed
- there should be no relationship between the residuals and the predicted value

- Residual is the value of the observed cost (actual value) - predicted value from equation

- R^2 measures the fit of the model to the data
97.8% of the variance of cost is accounted for by the model

c. 2.82 + 0.045t

t = 25 = 2.82 + 0.045(25) = 3.945%
Comparing this to the scatter plot the answer is in line with the data

t = 50 = 2.82 + 0.045(50) = 5.07
- this t value is outside of the storage time given but is an accurate result.
- Contrary to most people's opinion it is not double the result for t = 25.