Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

0. Chi square statistic.
1. Review list.
2. Review formulas.
3. Review exercises.

The final is Fri Dec14, 8-11am.
Bring a PENCIL and CALCULATOR and any books or notes you want. No computers.
http://www.stat.ucla.edu/~frederic/13/F18.

Chi square statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$

- Another statistic commonly used is the chi-square statistic.
- Its distribution is easier to fit with a theory-based distribution than a MAD statistic distribution.

$$\chi^2 = \left(\frac{1}{\hat{p}(1-\hat{p})} \right) \sum n_i (\hat{p}_i - \hat{p})^2$$

- $\hat{p}_i$ is the proportion of successes in category *i*
- $\hat{p}$ is the overall proportion of successes in the dataset.
- n_i is the sample size of category *i*

Statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- Another statistic commonly used is the chi-square statistic.
- Its distribution is easier to fit with a theory-based distribution than a MAD statistic distribution.

$$\chi^2 = \left(\frac{1}{\hat{p}(1-\hat{p})} \right) \sum n_i (\hat{p}_i - \hat{p})^2$$

- $\hat{p}_i$ is the proportion of successes in category *i*
- $\hat{p}$ is the overall proportion of successes in the dataset.
- n_i is the sample size of category *i*

1. Review list.

1. Meaning of SD.
2. Parameters and statistics.
3. Z statistic for proportions.
4. Simulation and meaning of pvalues.
5. SE for proportions and means.
6. What influences pvalues.
7. CLT and validity conditions for tests.
8. 1-sided and 2-sided tests.
9. Reject the null vs. accept the alternative.
10. Sampling and bias.
11. Significance level.
12. Type I, type II errors, and power.
13. CIs for a proportion.
14. CIs for a mean.
15. Margin of error.
16. Practical significance.
17. Confounding.
18. Observational studies and experiments.
19. Random sampling and random assignment.
20. Two proportion CIs and testing.
21. Median, IQR, and boxplots.
22. CIs for 2 means and testing.
23. Paired data.
24. Placebo effect, adherer bias, and nonresponse bias.
25. Prediction and causation.
26. Multiple testing and publication bias
27. Polling errors.
28. Correlation.
29. Regression.
30. Calculate & interpret a & b.
31. Goodness of fit in regression, resid. plot.
32. Common regression problems (causation, extrapolation,curvature, heteroskedasticity).
33. Comparing multiple means with MAD.
34. ANOVA & F-test.
35. Comparing multiple percentages using simulations and chi square test.

2. Review formulas.

1 proportion. $\hat{p}$ = proportion in sample with the property, π = pop. proportion with property, SE = √($\hat{p}$ (1-$\hat{p}$)/n). Z statistic = ($\hat{p}$-π)/SE.
There should be at least 10 of each type in the sample for the Z-test to be valid.
For a 95% CI, use $\hat{p}$ +/- 1.96 SE.

1 mean, t-test. xbar = sample mean, μ = population mean, s = sample sd, σ = pop. sd.
SE of mean = σ/√n, estimated using s/√n if σ is unknown.
t = (xbar – μ0)/(SE of mean), where μ0 = pop. mean under the null.
The population should be normally distributed for the t-test to be valid.
If the sample size n is large (> 30), or the pop. is normal and σ is known, then can do a Z-test, where Z = (xbar – μ0)/(SE of mean).

2 proportions. $\hat{p}1$ = proportion of treatment group, $\hat{p}2$ = proportion of control group.
$\hat{p}_{all}$ = pooled proportion = total # in both groups with the property / (n1+n2).
SE for difference = $\sqrt{[\hat{p}_{all}(1-\hat{p}_{all})(1/n1 + 1/n2)]}$. z = ($\hat{p}1$-$\hat{p}2$)/SE for difference.
For CIs, the SE for difference = $\sqrt{[\hat{p}1(1-\hat{p}1)/n1 + \hat{p}2(1-\hat{p}2)/n2]}$.
For a 95% CI use $\hat{p}1$-$\hat{p}2$ +/- 1.96 SE for difference.

2. Review formulas continued.

2 means. x1bar = sample mean 1, x2bar = sample mean 2, s_1 = sd of group 1, s_2 = sd 2.
SE for difference = $\sqrt{(s_1^2/n1+s_2^2/n2)}$. t = (x1bar-x2bar)/SE for difference.
The conditions are basically the same as for 1 mean.

For paired data on 2 means or 2 percentages, consider the differences for each person and treat it as 1 mean or 1 proportion.
But for simulation, randomly multiply each difference by either 1 or -1.

Regression and correlation. $\hat{y}$ = a + bX, where $\hat{y}$ means the prediction of Y using regression.
r is the sample correlation between X and Y. ρ is the pop. correlation.
Slope b = r s_y/s_x. Intercept a = ybar - b xbar.
To test if the slope, or equivalently if the correlation, is significantly different from 0,
SE of correlation = $\sqrt{[(1-r^2)/(n-2)]}$. t = r/SE of correlation.
For a t-test the data should be roughly elliptical, i.e. both X and Y should be roughly normal.
r^2 = proportion of variance in Y explained by the regression line.
$[\sqrt{(1-r^2)}]s_y$ = residual standard error = SD of residuals = how much regression predictions would typically be off by.

2. Review formulas continued.

Comparing 3 or more groups.
mad = average of absolute differences between group means.
For 3 groups for instance, mad = 1/3(|x1bar-x2bar|+|x1bar-x3bar|+|x2bar-x3bar|).
F = variability between groups / variability within groups = $MS_{treatment}$ / MS_{error} using ANOVA table. For F-test, draws should be normal and no group SD should be more than double any other group SD.

$$\chi^2 = \left(\frac{1}{\hat{p}(1-\hat{p})}\right)\sum n_i(\hat{p}_i - \hat{p})^2$$

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. $b = r\, s_y/s_x = 0.20\ (13)/(4) = 0.65$ years/inch.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

b. a = ybar – bxbar = 82 – 0.65(66) = 39.1 years.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

c. 39.1 + 0.65(68) = 83.3 years.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

d. $[\sqrt{}(1-r^2)]\ sy = 13\ \sqrt{}(1-.2^2) = 12.7$ years.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. b = r sy/sx = 0.20 (13)/(4) = 0.65 years/inch.
b. a = ybar – bxbar = 82 – 0.65(66) = 39.1 years.
c. 39.1 + 0.65(68) = 83.3 years.
d. $[\sqrt{(1-r^2)}]$ sy = 13 $\sqrt{(1-.2^2)}$ = 12.7 years.
e. r^2 = 0.04.
f. No, this is an observational study so it might be prone to confounding factors. Perhaps those with poorer diet are more likely to be smaller and also more likely to die younger. Thus the diet is the cause, not height.
g. No, this would be extrapolation.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

e. $r^2 = 0.04$.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

f. No, this is an observational study so it might be prone to confounding factors. Perhaps those with poorer diet are more likely to be smaller and also more likely to die younger. Thus the diet is the cause, not height.

3. Review exercises.

1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

a. What is the slope of the regression line predicting longevity using height?
b. What is the intercept of the regression line predicting longevity using height?
c. What would be the predicted lifespan, using the regression line, for someone 68 inches tall?
d. How much would your prediction using the regression line typically be off by?
e. What proportion of the variation in longevity is explained by the regression line?
f. Can we conclude that the stress of being small causes people to die younger?
g. Can we use this regression line to make an accurate prediction for someone whose height is 89 inches?

3. Review exercises.

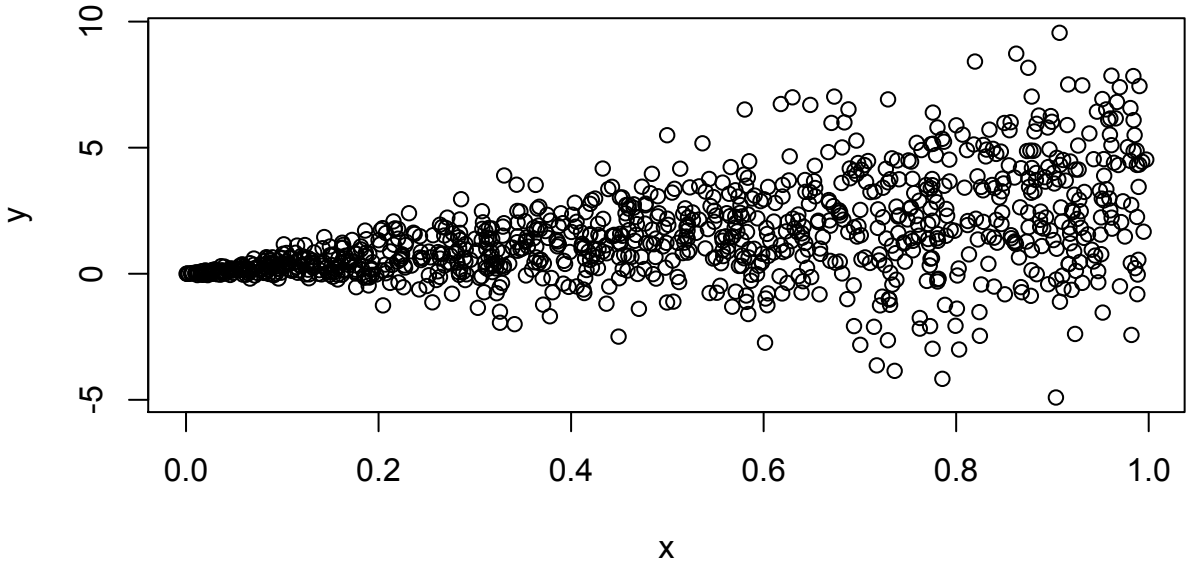
1. Suppose a researcher is trying to predict how long Americans live (Y) based on their adult height (X). She takes a simple random sample of 200 Americans and observes them until their death, and finds their mean height is 66 inches with a standard deviation of 4 inches. She finds their mean age at death is 82 years with a SD of 13 yrs. Both variables seem approximately normally distributed. The sample correlation between height and weight is 0.20.

g. No, this would be extrapolation. This person would be 5.75 SDs above the mean in height.

3. Review exercises.

2. Suppose a group of 100 bald men, 200 men with short hair, and 300 men with long hair are sampled and their pulses are measured. The mean pulse of the bald men is 67 bpm, the mean pulse of the short haired men is 70bpm, and the mean pulse of the long haired men is 72bpm. Calculate the mad for these 3 group means.

3. Which word describes a key feature of this plot?



4. Suppose the bald men described above have a mean pulse of 67 bpm and an sd of 10 bpm. If you select a bald man at random and guess his pulse is 67bpm, how much will you typically be off by?

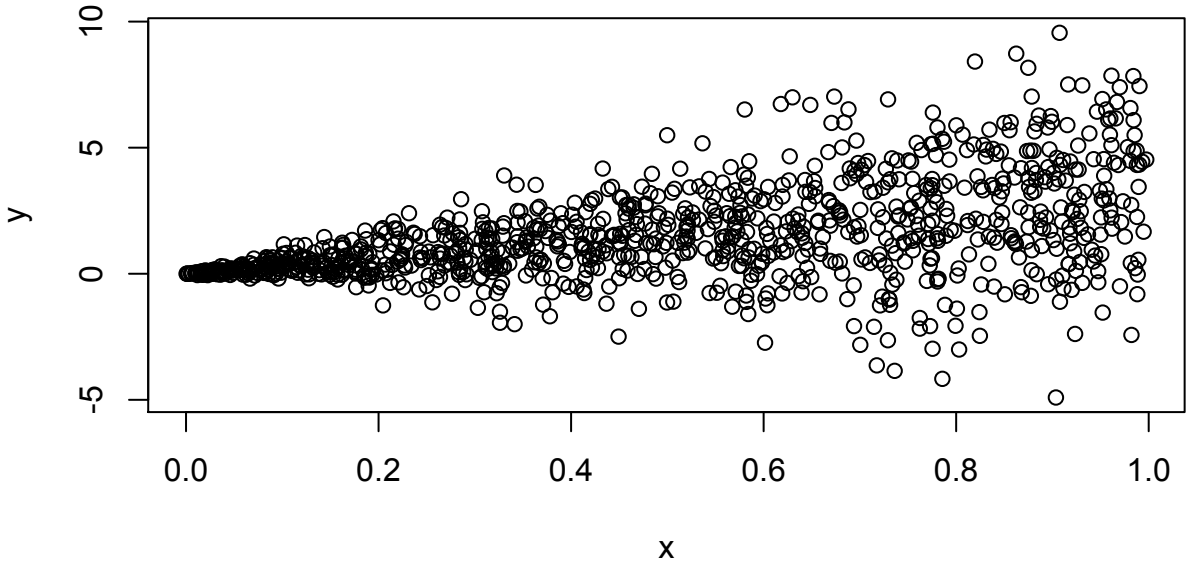
3. Review exercises.

2. 1/3(|67-70| + |67-72| + |70-72|) = 1/3(10) = 3.33.

3. Review exercises.

2. Suppose a group of 100 bald men, 200 men with short hair, and 300 men with long hair are sampled and their pulses are measured. The mean pulse of the bald men is 67 bpm, the mean pulse of the short haired men is 70bpm, and the mean pulse of the long haired men is 72bpm. Calculate the mad for these 3 group means.

3. Which word describes a key feature of this plot?



4. Suppose the bald men described above have a mean pulse of 67 bpm and an sd of 10 bpm. If you select a bald man at random and guess his pulse is 67bpm, how much will you typically be off by?

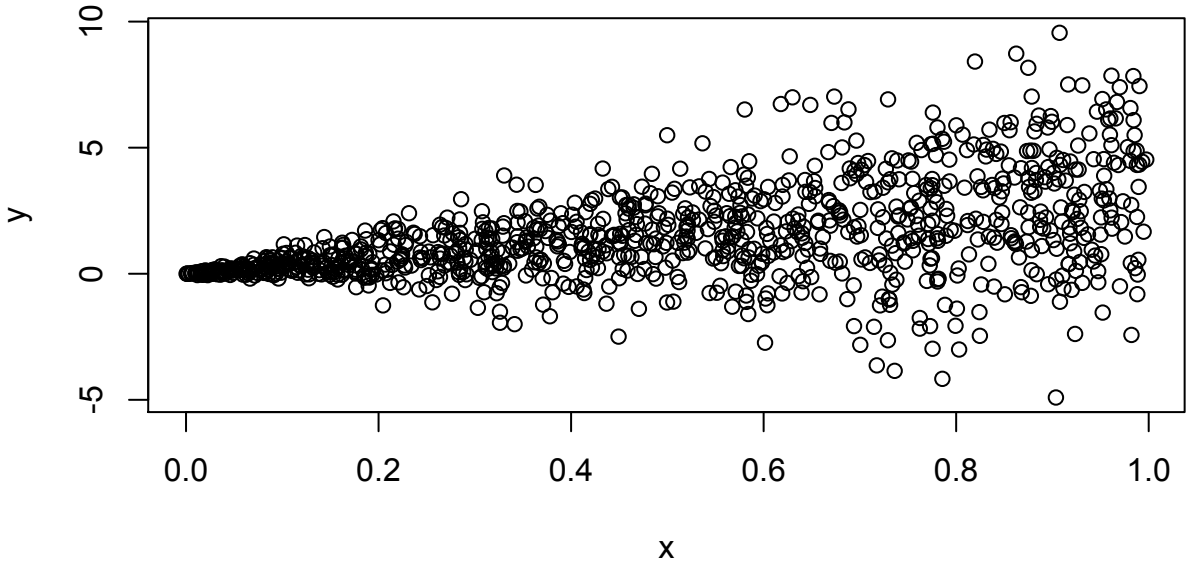
3. Review exercises.

3. Heteroskedasticity.

3. Review exercises.

2. Suppose a group of 100 bald men, 200 men with short hair, and 300 men with long hair are sampled and their pulses are measured. The mean pulse of the bald men is 67 bpm, the mean pulse of the short haired men is 70bpm, and the mean pulse of the long haired men is 72bpm. Calculate the mad for these 3 group means.

3. Which word describes a key feature of this plot?



4. Suppose the bald men described above have a mean pulse of 67 bpm and an sd of 10 bpm. If you select a bald man at random and guess his pulse is 67bpm, how much will you typically be off by?

3. Review exercises.

4. 10 bpm.

3. Review exercises.

5. Suppose the 100 bald men with a mean pulse of 67bpm have a median pulse of 70bpm. Describe the distribution of the pulses of bald men.

a. Left-skewed.

b. Right-skewed.

c. Symmetric.

d. Normal.

e. Confounding factors with heteroskedastic t-test confidence intervals.

f. None of the above.

6. Suppose the IQR is 20bpm, the range is (50bpm, 100bpm), and the 75^{th} percentile is 80bpm. What is the 25^{th} percentile pulse?

7. Find a 95% CI for the difference in mean pulse between the bald men and the long haired men. Assume the 100 bald men have mean 67bpm and sd 10bpm, and the 300 long haired men have mean 72bpm and sd 14bpm.

8. What can we conclude about statistical significance, based on this 95% CI?

3. Review exercises.

5. Left skewed, because the mean < median.

3. Review exercises.

5. Suppose the 100 bald men with a mean pulse of 67bpm have a median pulse of 70bpm. Describe the distribution of the pulses of bald men.

a. Left-skewed.

b. Right-skewed.

c. Symmetric.

d. Normal.

e. Confounding factors with heteroskedastic t-test confidence intervals.

f. None of the above.

6. Suppose the IQR is 20bpm, the range is (50bpm, 100bpm), and the 75^{th} percentile is 80bpm. What is the 25^{th} percentile pulse?

7. Find a 95% CI for the difference in mean pulse between the bald men and the long haired men. Assume the 100 bald men have mean 67bpm and sd 10bpm, and the 300 long haired men have mean 72bpm and sd 14bpm.

8. What can we conclude about statistical significance, based on this 95% CI?

3. Review exercises.

6. 60bpm.

3. Review exercises.

5. Suppose the 100 bald men with a mean pulse of 67bpm have a median pulse of 70bpm. Describe the distribution of the pulses of bald men.

a. Left-skewed.

b. Right-skewed.

c. Symmetric.

d. Normal.

e. Confounding factors with heteroskedastic t-test confidence intervals.

f. None of the above.

6. Suppose the IQR is 20bpm, the range is (50bpm, 100bpm), and the 75^{th} percentile is 80bpm. What is the 25^{th} percentile pulse?

7. Find a 95% CI for the difference in mean pulse between the bald men and the long haired men. Assume the 100 bald men have mean 67bpm and sd 10bpm, and the 300 long haired men have mean 72bpm and sd 14bpm.

8. What can we conclude about statistical significance, based on this 95% CI?

3. Review exercises.

7. -5 +/- 1.96 SE, where SE = $\sqrt{(100/100 + 14^2/300)} = 1.28582$. So the 95% CI is -5 +/- 1.96(1.28582) = -5 +/- 2.53.

3. Review exercises.

5. Suppose the 100 bald men with a mean pulse of 67bpm have a median pulse of 70bpm. Describe the distribution of the pulses of bald men.

a. Left-skewed.

b. Right-skewed.

c. Symmetric.

d. Normal.

e. Confounding factors with heteroskedastic t-test confidence intervals.

f. None of the above.

6. Suppose the IQR is 20bpm, the range is (50bpm, 100bpm), and the 75^{th} percentile is 80bpm. What is the 25^{th} percentile pulse?

7. Find a 95% CI for the difference in mean pulse between the bald men and the long haired men. Assume the 100 bald men have mean 67bpm and sd 10bpm, and the 300 long haired men have mean 72bpm and sd 14bpm.

8. What can we conclude about statistical significance, based on this 95% CI?

3. Review exercises.

8. 0 is not in the interval, so the difference in pulse between bald men and long haired men is statistically significant.