Stat 430 Homework 2

due: October 16, 2009 at 5pm

1. Access the reading scores data. These data are the reading scores of grade school students after they received a directed reading activities ("Treated") or not ("Control"). Construct 99% confidence interval for the difference in population means between treated and control. Report the p-value for testing the null hypothesis of equal population means. What is the power to detect a score difference of 5?

Solution: The simplest way is to use R:

```
d <- read.table("http://thirteen-01.stat.iastate.edu/wiki//stat430/files?filename=read.txt", header=T)
t.test(d$Response[d$Treatment=="Treated"], d$Response[d$Treatment=="Control"], conf.level=0.99)
```

with output

```
Welch Two Sample t-test

data:  d$Response[d$Treatment == "Treated"] and d$Response[d$Treatment == "Control"]
t = 2.3109, df = 37.855, p-value = 0.02638
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -1.728272 21.637175
sample estimates:
mean of x mean of y
 51.47619  41.52174
```

The confidence interval is $(-1.73, 21.64)$, and the p-value is $0.026 < 0.05$, and we reject H_0 :treatment has no impact on reading scores (but note, there is no rejection if we insist to control type I error at or below $\alpha = 0.01$, as was used for CI).

For the power calculation, some of you found R's `power.t.test()` function, but it is necessary to provide an estimate of the sampling variance, which is assumed to be the same for the two samples. We provide that estimate by assuming common variance and computing the pooled variance.

```
var.treat <- var(d$Response[d$Treatment=="Treated"])
var.control <- var(d$Response[d$Treatment=="Control"])
n.treat <- sum(d$Treatment=="Treated")
n.control <- sum(d$Treatment=="Control")
var.pool <- ((n.treat-1)*var.treat + (n.control-1)*var.control)/(n.treat+n.control-2)
power.t.test(n=c(21,23), delta=5, sd=sqrt(var.pool), type="two.sample")
     Two-sample t test power calculation

              n = 21, 23
          delta = 5
             sd = 14.55120
      sig.level = 0.05
          power = 0.1912981, 0.2060974
    alternative = two.sided

 NOTE: n is number in *each* group
```

The power is reported to be between 0.19 and 0.21 if the common sample size is between $n = 21$ and $n = 23$.

Let's examine the previous analysis a little more carefully. By default, `R` performs a Welch Two Sample t-test, where the variances of the two samples are assumed to be unequal. We can check this assumption, roughly, by calculating the variances of both samples (with output interspersed):

```
var(d$Response[d$Treatment=="Treated"])
[1] 121.1619   # sd = 11.00736, mean = 51.47619
var(d$Response[d$Treatment=="Control"])
[1] 294.0791   # sd = 17.14873, mean = 41.52174
```

Indeed, they do look different, though at this time we had not learned any method to test for significance of this difference.

In any case, we are hesitant to assume common variance, so the Welch test seems appropriate. To implement it by hand, we perform the following calculations (with output interspersed):

```
var.diff <- var.treat/n.treat + var.control/n.control
df <- (var.treat/n.treat + var.control/n.control)^2/((var.treat/n.treat)^2/n.treat
   + (var.control/n.control)^2/n.control)-2
mean.treat <- mean(d$Response[d$Treatment=="Treated"])
mean.control <- mean(d$Response[d$Treatment=="Control"])
t <- (mean.treat-mean.control)/sqrt(var.diff)
2*pt(-abs(t), df=df)
[1] 0.02642018
```

The 99% confidence intervals are computed as (with output interspersed and corrected from first posting)

```
mean.treat-mean.control - qt(0.995, df=df)*sqrt(var.diff)
[1] -1.732332
mean.treat-mean.control + qt(0.995, df=df)*sqrt(var.diff)
[1] 21.64123
```

The power calculation, assuming that the variance we computed of the difference $\bar{X} - \bar{Y}$ is the truth and using the exact sample sizes observed, is

```
> 1-pnorm(qnorm(0.975) - 5/sqrt(var.diff)) + pnorm(qnorm(0.025) - 5/sqrt(var.diff))
[1] 0.2129800
```

2. For independent samples

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_X, \sigma_X^2)$$
$$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_Y, \sigma_Y^2)$$

show that $\bar{X} - \bar{Y}$ is the maximum likelihood estimate for $\mu_X - \mu_Y$.

Solution: We know

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2 + \sigma_Y^2}{n}\right)$$

so the likelihood is

$$L(\mu_X - \mu_Y, \sigma_X^2, \sigma_Y^2) = \frac{1}{\sqrt{2\pi \frac{\sigma_X^2+\sigma_Y^2}{n}}} \exp\left[-\frac{\left[\bar{X} - \bar{Y} - (\mu_X - \mu_Y)\right]^2}{\frac{2(\sigma_X^2+\sigma_Y^2)}{n}}\right]$$

with log likelihood

$$l(\mu_X - \mu_Y, \sigma_X^2, \sigma_Y^2) \propto -\frac{\left[\bar{X} - \bar{Y} - (\mu_X - \mu_Y)\right]^2}{\frac{2(\sigma_X^2+\sigma_Y^2)}{n}}$$

after dropping terms not involving $\mu_X - \mu_Y$. The equation to solve is obtained by taking the derivative with respect to $\mu_X - \mu_Y$ and setting the result to 0.

$$\frac{dl(\mu_X - \mu_Y, \sigma_X^2, \sigma_Y^2)}{d(\mu_X - \mu_Y)} = \frac{2\left[\bar{X} - \bar{Y} - (\mu_X - \mu_Y)\right]}{\frac{2(\sigma_X^2+\sigma_Y^2)}{n}} = 0$$

which yields

$$\widehat{\mu_X - \mu_Y} = \bar{X} - \bar{Y}.$$

There is a long way, via the joint distribution of $X_1, \ldots, X_n, Y_1, \ldots, Y_n$, but it requires substantially more algebra to result in the same likelihood $L(\mu_X - \mu_Y, \sigma_X^2, \sigma_Y^2)$ above that we know from reproductive property of normally distributed random variables.

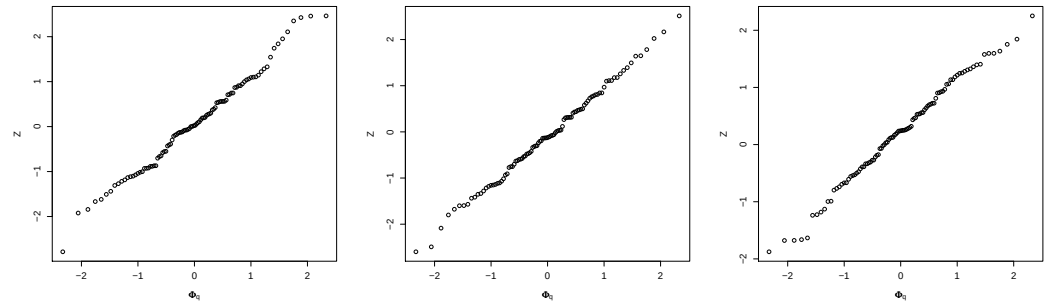
3. In this problem, you will explore the shape of the probability plot for various normal and non-normal distributions and additional tests for normality.

   (a) Simulate data sets of 100 standard normal random using `R`'s `rnorm()` function. Prepare probability plots (x-axis: standard normal quantiles; y-axis: observed and ordered random variables) for three such data sets.

   Solution: The following `R` code produces the plots below. These plots are supposed to show points falling along a straight line. Deviations from a straight line, particularly noticeable near the tails, indicate how much variation is typical in such plots even when the data *are* normally distributed.

```
n <- 100
for(i in 1:3) {
   x <- rnorm(n=n)
   plot(qnorm((1:n)/(n+1)), sort(x), type="p", cex.lab=1.2, cex.axis=1.2, main="", ylab="Z",
```

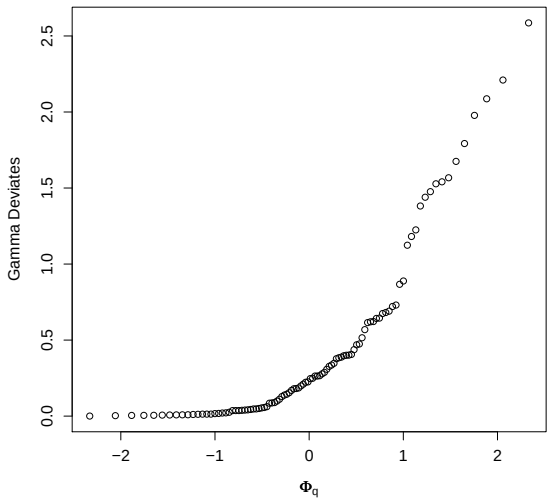
```
        xlab=expression(Phi[q]))
}
```



(b) Simulate 100 random variables from a Gamma(shape= 0.471, rate= 1) distribution using R's `rgamma()` function. Prepare a probability plot (x-axis: standard normal quantiles; y-axis: ordered gamma-distributed random variables). Based on this plot, how does this distribution differ from the normal? Save this data for part d.

Solution: The following R code produces the probability plot below. The upward curvature is quite clearly indicative of a contracted left tail and an extended right tail, relative to the standard normal plot (see further discussion of how to interpret tails in part c). The curvature is substantially more obvious in this plot than in any of the plots of part a.

```
d.gamma <- rgamma(n=n, shape=0.471, rate=1)
plot(qnorm((1:n)/(n+1)), sort(d.gamma), type="p", cex.lab=1.2, cex.axis=1.2, main="",
   ylab="Gamma Deviates", xlab=expression(Phi[q]))
```

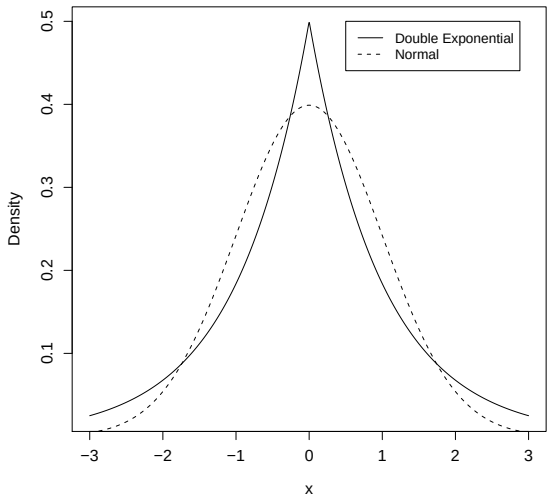


(c) The double exponential distribution has pdf

$$f(x) = \frac{1}{2}e^{-|x|} \qquad\qquad -\infty < x < \infty.$$

Compare this pdf with the normal pdf. Do the tails of this distribution decay slower or faster? Generate 100 random variables from the double exponential distribution by (1) finding its cdf $F(x)$, (2) deriving the inverse cdf $F^{-1}(z)$, (3) generating 100 standard uniform random variables $Z_1, \ldots, Z_{100}$ using R's `runif()` function, and (4) mapping them to exponential random variable $X_1 = F^{-1}(Z_1), \ldots, X_{100} = F^{-1}(Z_{100})$ using the inverse cdf. Save this data for part d. Prepare a probability plot. Does it confirm your suspicions about the tails?

Solution: Below is a plot of the two distributions in question. Clearly, the tails of the double exponential decay slower than the tails of the normal distribution.



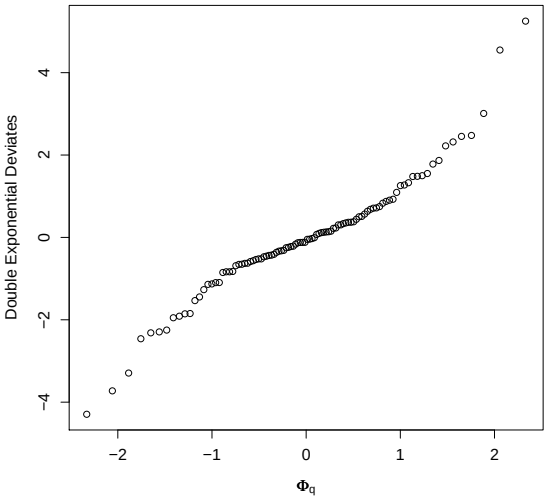
The cdf is

$$F(x) = \begin{cases} \frac{1}{2}e^{x} & x \leq 0 \\ 1 - \frac{1}{2}e^{-x} & x > 0 \end{cases}$$

If we let $z = F(x)$, then

$$F^{-1}(z) = \begin{cases} \ln(2z) & z \leq 1/2 \\ -\ln[2(1-z)] & z > 1/2 \end{cases}$$

The probability plot for $n = 100$ random double exponential deviates is shown below. The sudden dip (on left) and rise (on right) of the points indicates that the quantiles of the double exponential are stretched away from the center (0) more than the quantiles of the standard normal. So, yes, the plot confirms the previous observation. However, the differences are more subtle for double exponential deviates than gamma deviates, though the curvature still seems more substantial than that observed in part a. Still, these are plots based on a sample size of $n = 100$, and the curvature will be less obvious for smaller data sets.



(d) The coefficient of skewness (a measure of skew) is estimated as

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{ns^3}$$

where s^2 is the sample variance. For normal random variables, this coefficient should be 0. Estimate this coefficient for the gamma-distributed random variables from part b and the double exponential-distributed random variables from part c. The question is whether either of these statistics is unusually different from 0.

A theoretical sampling distribution for b_1 is difficult to obtain, but the computer can help you by allowing you to sample coefficients of skewness under the assumption of $n = 100$ iid normal samples. To do this, generate $n = 100$ iid standard normal random variables, and compute the coefficient of skewness. Repeat $B = 1000$ times to obtain bootstrap skewness coefficients $b_1^{(1)}, \ldots, b_1^{(1000)}$. The fraction of bootstrap coefficients that are as extreme or more extreme than the coefficient computed on the original data is the p-value for rejecting the null hypothesis. Test for skewness of the gamma and double exponential random variables using this test.

Repeat the above procedure for the coefficient of kurtosis (a measure of peakedness or flatness), which is estimated as

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{ns^4}$$

Solution: The code for doing the calculations includes two functions that compute skewness and kurtosis.

```
skew <- function(x) {
   m <- mean(x)
```

```
   s <- sd(x)
   n <- length(x)
   sum((x-m)^3/n/s^3)
}

kurtosis <- function(x) {
   m <- mean(x)
   s <- sd(x)
   n <- length(x)
   sum((x-m)^4/n/s^4)
}
B <- 1000
# simulate standard normal data
X <- apply(matrix(n, nrow=B, ncol=1), 1, rnorm)
# compute skewness and kurtosis for standard normal
S <- apply(X, 2, skew)
K <- apply(X, 2, kurtosis)
# compute skewness/kurtosis for gamma and double exponential data sets
S.gamma <- skew(d.gamma)
S.dexp <- skew(d.dexp)
K.gamma <- kurtosis(d.gamma)
K.dexp <- kurtosis(d.dexp)
# compute fraction of bootstraps as extreme or more extreme than observed
p.gamma.S <- sum(abs(S.gamma) <= abs(S))/B
p.dexp.S <- sum(abs(S.dexp) <= abs(S))/B
p.gamma.K <- sum(abs(K.gamma) <= abs(K))/B
p.dexp.K <- sum(abs(K.dexp) <= abs(K))/B
# output results
cat("Tests for gamma\n\tobserved skew", S.gamma, "; skewness p-value", p.gamma.S,
   "\n\tobserved kurtosis", K.gamma, "; kurtosis p-value", p.gamma.K, "\n")
cat("Tests for double exponential\n\tobserved skew", S.dexp, "; skewness p-value", p.dexp.S,
   "\n\tobserved kurtosis", K.dexp, "; kurtosis p-value", p.dexp.K, "\n")
```

The output is below. The skew is significantly non-zero for the gamma distribution only. The results makes sense. The gamma distribution is skewed, the double exponential is symmetric. Kurtosis is significantly non-zero for both the gamma and double exponential distribution. This results makes sense particularly for the double exponential, which clearly differs in peakedness.

```
Tests for gamma
   observed skew 1.709914 ; skewness p-value 0
   observed kurtosis 5.284533 ; kurtosis p-value 0.002
Tests for double exponential
   observed skew 0.3489875 ; skewness p-value 0.121
   observed kurtosis 5.237847 ; kurtosis p-value 0.002
```

4. Access the Titanic data to answer the following questions.

   (a) Use a t-test to test whether there is a difference in the mean age of people who survived and did not survive the disaster. Is there a difference in the mean age of people in first and second class (variable "PClass")?

Solution: Let μ_S be the mean age of survivors and μ_D the mean age of non-survivors, then a t-test of

$$H_0 : \mu_S = \mu_D$$

uses statistic $t = -1.99$ with p-value $p = 0.04655$, which provides just barely "legal" evidence that there is a difference in the ages.

Let μ_U be the mean age of people in first and second class, while μ_L is the mean age of people in the lower class. Then, a t-test of

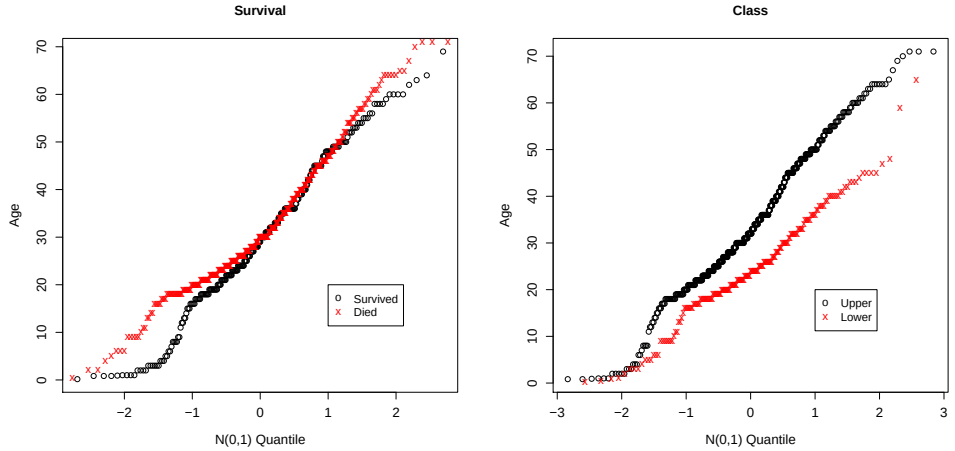
$$H_0 : \mu_U = \mu_L$$

uses statistic $t = 8.877$ with very significant p-value.

For both tests, we do not assume common variance and apply a Welch test.

(b) Prepare probability plots to evaluate the appropriateness of the normal assumption for the two tests above. Which test results are most suspect? Can you think of a transformation that might fix this problem? (No need to implement, just discuss.)

Solution:



Keeping in mind that the data set size is quite large, there are some non-trivial patterns in the data. These patterns seem to vary with the subsample examined. For example, age usually has a gamma-type truncated left tail, which makes sense when one recognizes that negative ages are not allowed. For the non-survivors, this pattern disappears. The upper tail of ages in upper class appears truncated, while it is elongated in lower class. Not being very creative, I can think of no transformation that does the job, certainly none of the ones we discussed in class and not universally for all four samples.

(c) Perform a test of whether the proportion of children two years and under among the survivors is different from among the non-survivors.

Solution: Let p_B and p_O be the probabilities of survival of young children (babies) and everyone else (others). We seek to test

$$H_0 : p_B = p_O$$

We estimate $\hat{p}_B = 0.8125$ and $\hat{p}_O = 0.434$. There certainly appears to be a difference. Under the null hypothesis, there is a universal proportion of survival $\hat{p} = \frac{n_B\hat{p}_B + n_O\hat{p}_O}{n_B+n_O} = 0.44$, where n_B is the number of babies and n_O is the number of others. The test statistic is $t = \frac{\hat{p}_B - \hat{p}_O}{\sqrt{\hat{p}(1-\hat{p})(1/n_B+1/n_O)}} = 3.01$. The p-value is 0.002650478, which suggests a significant difference in survival rate of babies. It is not surprising that babies were preferentially saved. In fact, you might have had so much confidence in human dignity regarding children, that you could have performed a one-sided test.

Another approach to this problem is to perform a Fisher's exact test, which has the added advantage of not requiring the CLT to obtain the p-value. The requisite `R` code and output is shown below, confirming our results above, but more dramatically.

```
d <- read.csv("titanice.csv", header=T)
d$less.2 <- NA
d$less.2[!is.na(d$age) & d$age<=2] <- T
d$less.2[!is.na(d$age) & d$age>2] <- F
fisher.test(table(d[,c("less.2","survived")]))

Fisher's Exact Test for Count Data

data:  table(d[, c("less.2", "survived")])
p-value = 0.003705
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.525652 31.108586
sample estimates:
odds ratio
  5.629998
```

(d) When data normality is suspect, the computer can help perform tests without distributional assumptions. For example, if there are no differences between the surviving and non-surviving populations, the combined age data can be viewed as a sample from a single population that is arbitrarily split into survivors and non-survivors. Use `R`'s `sample()` function to randomly sample 1000 of these assignments. For each one, compute a statistic sensitive to differences in population means. Count the number of times (out of 1000) where the statistic is as extreme or more extreme than the statistic computed on the original data. Divide this count by 1000 to obtain a p-value for testing the null hypothesis. How do the assumptions of this test differ from the assumptions of the two-sample t-test?

Solution: The R code is shown below:

```
age <- d$age[!is.na(d$age)]   # extract age data only (discard passengers with no known age)
n.with.age <- length(age)     # number of passengers with age
n.survivors <- sum(d$survived==1 & !is.na(d$age))   # number of survivors with age
# compute observed statistic
T.obs <- (mean(d$age[d$survived==1], na.rm=T)
   - mean(d$age[d$survived==0], na.rm=T))/sqrt(var(d$age[d$survived==1], na.rm=T)/n.survivors
   + var(d$age[d$survived==0], na.rm=T)/(n.with.age-n.survivors))
cnt <- 0
for(i in 1:1000) {
   age.permute <- sample(age, replace=F)   # permute ages as if survivorship didn't matter
   # compute statistic on permuted data
   T <- (mean(age.permute[1:n.survivors]) - mean(age.permute[(n.survivors+1):n.with.age]))
      /sqrt(var(age.permute[1:n.survivors])/n.survivors
         + var(age.permute[(n.survivors+1):n.with.age])/(n.with.age-n.survivors))
   # count if permuted data is as or more extreme than observed
   if(abs(T) >= abs(T.obs)) cnt <- cnt + 1
}
```

The p-value is the variable `cnt` (the number of times the simulated statistic is as extreme or more extreme than the observed statistic) divided by the number of permutations, which in my case, turned out to be 0.049, just *barely* significant.

(e) [added] Using the extended Titanic data, pair each survivor with a random non-survivor with the same embarkation location (consider two missing embarkation locations as matched). You will need to discard several non-survivors that have no match. Is the p-value for rejecting the null hypothesis of no age difference smaller or larger? Propose an explanation for what you find.

Solution: Please see the R code provided elsewhere (warning: it may not provide the most elegant solution for this part). The new p-value will vary depending how you setup the matches. One iteration of random matching in my code lead to p-value 0.003473123. In fact, most matchings gave significantly smaller p-values. Despite throwing out 133 unmatched individuals, the result is more significant.

(f) [added] Use one of the non-parametric tests we learned in class to test for a difference in mean age of survivors and non-survivors.

Solution: The test that makes sense (and no other tests really makes sense unless you use the paired data from the last part) is the Wilcoxon rank sum test. Using R for this yields (with output)

```
wilcox.test(d$age[d$survived==1 & !is.na(d$age)], d$age[d$survived==0 & !is.na(d$age)])
Wilcoxon rank sum test with continuity correction

data:  d$age[d$survived == 1 & !is.na(d$age)] and d$age[d$survived == 0 & !is.na(d$age)]
W = 45890, p-value = 0.1187
alternative hypothesis: true location shift is not equal to 0
```

Shock! No significant difference.

(g) [added] Questions a through f all address the question of whether age affected

passenger survival. What is your overall conclusion? Why?

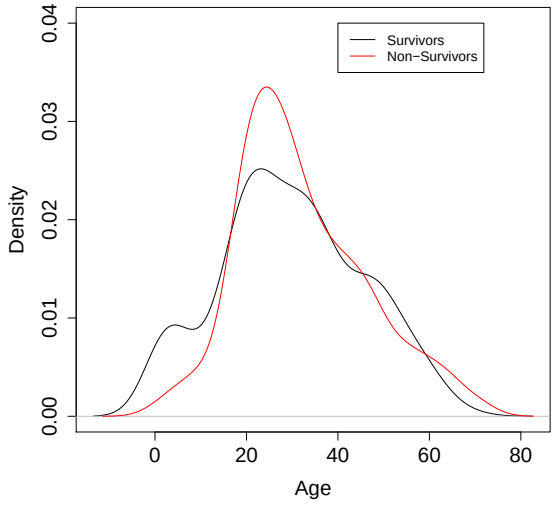
Solution: Using the two-sample t-test, there was borderline evidence of a mean age difference between survivors and non-survivors. The evidence of non-normality (in the probability plots) might make the results suspect, but the sample size (281 survivors and 352 non-survivors) should be large enough for CLT asymptotics.

The test of proportions yielded a more significant result, probably because it focuses on the main difference in the age distributions, namely the number of 0-year old survivors. However, the test of proportions uses CLT arguments and the sample size for the number of babies is only 16, perhaps not enough to assume CLT asymptotics. (Thus, Fisher's exact test would be a good choice, but you were not asked for it.)

The permutation test yielded a barely significant result, like the two-sample t-test. It is not surprising that these two tests agree, as they test the same thing and should give the same result when the CTL asymptotics are valid.

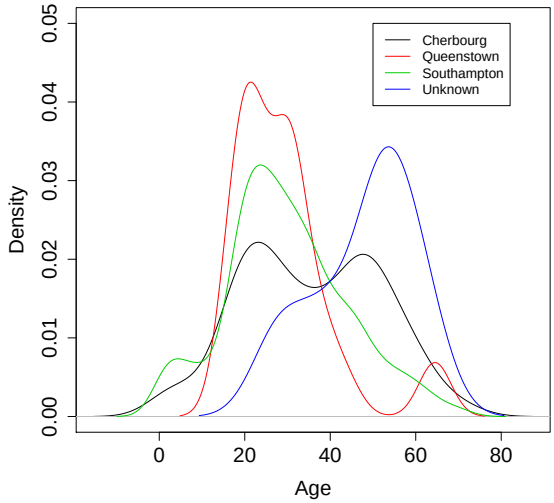
The rank sum test did not yield a significant result, probably because the test has such low power.

All of the above tests, except the test of proportions, is testing for a difference in the center of the age distribution. Since the main difference appears to be subtle difference in the left tail rather than a wholesale shift of the distribution (see estimated density plots below), all these tests may be less powerful than one would hope.



The matched pair t-test yielded the most significant result. The sample size is still large enough ($n = 250$) to hope that asymptotics still apply. The more significant result probably reflects a difference in age distributions of the people boarding at the distinct ports (see some confirmation below). Once this variabil-

ity in age distribution is removed (by taking differences), the difference between population means of survivors and non-survivors becomes more obvious; power has increased.



Overall, we conclude that there is a difference in mean age of survivors vs. non-survivors. It is nearly buried in the noise, but can be made more notable by matching on embarkation port.

(h) [added] Use a Fisher's exact test to test whether survival is related to sex and, separately, class (combining 1st and 2nd class together).

Solution: R code with output interspersed is

```
fisher.test(table(d[,c("sex","survived")]))
Fisher's Exact Test for Count Data

data:  table(d[, c("sex", "survived")])
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0775930 0.1338484
sample estimates:
odds ratio
 0.1021212
fisher.test(table(d[,c("merged.pclass","survived")]))
Fisher's Exact Test for Count Data

data:  table(d[, c("merged.pclass", "survived")])
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 3.501428 5.809690
sample estimates:
odds ratio
  4.501561
```

Both yield significant results showing there is a significant difference in the sex

of survivors (more females survived) and the class of survivors (more people in lower class died). The evidence for the direction of the conclusions is shown in the tables below.

Sex	Died	Survived
female	156	307
male	708	142

Class	Died	Survived
lower	574	137
upper	290	312

5. Form teams of 1 to 4 members and prepare a first proposal of your class project. You should identify a dataset and questions you would like to address with this dataset. If you don't have good ideas for datasets, I may be able to provide you with one. In this case, let me know what type of statistical expertise you would like to develop in this project. Try to be at least a little specific.