

Summary Sheet for Continuous Outcomes (One Sample and Two Samples)

1. Graphical displays for single sample

(a) Stem-and-leaf plot

- Select leading digits for stem values and trailing digits for leaves.
- Sample R commands:

```
> stem(ifng.cb)
# scale=2 doubles the number of stems
> stem(ifng.cb,scale=2)
```

(b) Histogram

- Divide data into non-overlapping classes.
- Count number of obs in each class.
- Draw rectangles with area proportional to frequencies and base equal to class intervals.
- Sample R commands:

```
> hist(ifng.cb)
# breaks=n gives approx 20 equal-sized intervals
> hist(ifng.cb,breaks=20)
# user-defined break points
> brk <- c(0,25,50,75,100,150,200,250,300,400,600,800,1000,1300)
> hist(ifng.cb, breaks=brk)
# deciles as break points
> hist(ifng.cb,breaks=quantile(ifng.cb,0:10/10))
```

(c) Dotplot

- Display dot for each observation.
- Stack duplicate observations vertically.
- Sample R commands:

```
> stripchart(ifng.cb,method='stack')
# log='x' gives log scale for x-axis
> stripchart(ifng.cb,method='stack',log='x')
```

(d) Box (or box-and-whiskers) plot

- Central line indicates median.
- Box indicates "hinges" (nearly quartiles).
- Whiskers indicate largest and smallest observations with a distance of `range` times the box size of the box.
- Default for `range` is 1.5; `range = 0` gives min and max.
- Sample R commands:

```
# whiskers extend to min and max
boxplot(ifng.cb, range=0)

# effect of range parameter
boxplot(ifng.cb, range=0, horizontal=T)
boxplot(ifng.cb, horizontal=T) #default range=1.5
boxplot(ifng.cb, range=2, horizontal=T)
```

(e) Empirical cumulative distribution (cdf)

- Fraction of data smaller than or equal to x .
- Sample R commands:

```
n <- length(ifng.cb)
# type='s' plots a step function
plot(sort(ifng.cb), (1:n)/n, type='s', ylim=c(0,1))
plot(sort(ifng.cb), (1:n)/n, type='s', ylim=c(0,1), log='x')
```

(f) Normal scores plot

- Plot k^{th} smallest observation against expected value of k^{th} smallest observation out of n standard normal rv's.
- Expect to obtain a straight line for data from a normal distribution with any mean and standard deviation.
- Sample R commands:

```
qqnorm(ifng.cb)
# add reference line to plot
qqline(ifng.cb)
```

2. Summary statistics for single sample

(a) Measures of central tendency

i. Arithmetic mean

- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
- Sample R command:
> mean(ifng.cb)

ii. Median

- Middle value of the ordered data.
- Sample R command:
> median(ifng.cb)

iii. Geometric mean

- $\tilde{x} = (\prod_{i=1}^n x_i)^{1/n}$
- $\log \tilde{x} = \frac{1}{n} \sum_{i=1}^n \log x_i$
- Only valid for non-negative data.
- Sample R command:
> exp(mean(log(ifng.cb)))

iv. Mode

- Most common data value.
- More useful for discrete data (many ties).
- Sample R command:
> as.numeric(names(which.max(table(ifng.cb))))

(b) Measures of dispersion

i. Range

- Difference between largest obs and smallest obs.
- Sample R commands:

```
> range(ifng.cb)
> diff(range(ifng.cb))
```

ii. Interquartile range (IQR)

- Difference between 3rd quartile and 1st quartile.
- Sample R commands:

```
> quantile(ifng.cb, c(.25, .75))
> diff(quantile(ifng.cb, c(.25, .75)))
```

iii. Mean absolute deviation

- Equals $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$.
- Sample R command:

```
> mean(abs(ifng.cb - mean(ifng.cb)))
```

iv. Variance and standard deviation

- Less intuitive than mean absolute deviation.
- More desirable mathematical properties (normal dist, CLT).
- Variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.
- Standard deviation: $s = \sqrt{s^2}$.
- Sample R commands:

```
> var(ifng.cb)
> sd(ifng.cb)
```

v. Coefficient of variation

- Ratio of standard deviation to mean.
- Relative spread of data.
- Only useful for non-negative data with absolute zero.
- Sample R command:

```
> sd(ifng.cb) / mean(ifng.cb)
```

3. Graphical displays for two or more samples

(a) Parallel histograms

- Separate histogram for each group (side-by-side or top-to-bottom).
- Need common set of classes for each plot.
- Sample R commands:

```
> par(mfrow=c(2,1))
# user-defined breaks
> brk <- seq(0.6,3.2,length=10)
> hist(log10.ifng.cb,breaks=brk,freq=T)
> hist(log10.ifng.1yr,breaks=brk,freq=T,col='gray')

# deciles of pooled sample as breaks
> brk <- quantile(c(log10.ifng.cb,na.omit(log10.ifng.1yr)),0:10/10)
> hist(log10.ifng.cb, breaks=brk)
> hist(log10.ifng.1yr, breaks=brk, col='gray')
> par(mfrow=c(1,1))
```

(b) Parallel box plots

- Side-by-side box plots for each group.
- Sample R commands:

```
> boxplot(log10.ifng.cb,log10.ifng.1yr,horizontal=T)
> boxplot(log10.ifng.cb,log10.ifng.1yr,horizontal=T,
names=c('Cord Blood','Age 1 yr'),
xlab='Interferon-gamma (log10 scale)')
```

(c) Empirical cdfs

- Empirical cdf for each group in one plot.
- Sample R commands:

```
# plot empirical cdf for cord blood
> n0 <- length(log10.ifng.cb)
> plot(sort(log10.ifng.cb),(1:n0)/n0,type='s',ylim=c(0,1))
# add empirical cdf for 1 year
> n1 <- length(na.omit(log10.ifng.1yr))
> lines(sort(na.omit(log10.ifng.1yr)),(1:n1)/n1,type='s',col='gray')
```

4. Paired (one-sample) tests

(a) One-sample t -test.

- Parameter of interest: population mean (μ).
- Hypotheses:
 - $H_0 : \mu = \mu_0$ (typically $\mu_0 = 0$ for paired data).
 - $H_1 : \mu \neq \mu_0$.
- Test statistic: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.
- Assumptions: Independence and either normality or large sample (CLT).
- Sample R commands:

```
> t.test(diff.log10.ifng)
```

```
One Sample t-test
```

```
data: diff.log10.ifng
t = -9.5136, df = 279, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.4218295 -0.2771915
sample estimates:
mean of x
-0.3495105

# alternative code for paired t-test
> t.test(log10.ifng.cb, log10.ifng.lyr, paired=T)
```

(b) Wilcoxon signed rank test.

- Parameter of interest: population median (μ).
- Hypotheses:
 - $H_0 : \mu = \mu_0$ (typically $\mu_0 = 0$ for paired data).
 - $H_1 : \mu \neq \mu_0$.
- Test statistic: Sum of the ranks of $|x_i - \mu_0|$ for $x_i > \mu_0$.
- Assumptions: Independence and symmetric distribution.
- Sample R commands:

```
> wilcox.test(diff.log10.ifng, correct=F)
```

```
Wilcoxon signed rank test
```

```
data: diff.log10.ifng
V = 7680, p-value < 2.2e-16
alternative hypothesis: true mu is not equal to 0

# include confidence interval for median
> wilcox.test(diff.log10.ifng, conf.int=T, correct=F)
```

```
Wilcoxon signed rank test
```

```
data: diff.log10.ifng
V = 7680, p-value < 2.2e-16
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
 -0.4469522 -0.3004784
sample estimates:
(pseudo)median
 -0.3745449

# alternative code for paired data
> wilcox.test(log10.ifng.cb, log10.ifng.1yr, paired=T, correct=F)
```

(c) Sign test.

- Parameter of interest: population median (μ).
- Hypotheses:
 - $H_0 : \mu = \mu_0$ (typically $\mu_0 = 0$ for paired data).
 - $H_1 : \mu \neq \mu_0$.
- Test statistic: Number of obs less than μ_0 .
- Assumptions: Independence.
- Sample R commands:


```
> sign.test <- function(x, mu=0, conf.level=0.95) {
  bhigh <- sum(x>=mu)
  n <- length(x)
  pbinom(bhigh,n,1/2)

  blow <- sum(x<=mu)
  pbinom(blow,n,1/2)
  cat("number of x >= mu: ", bhigh, "\n")
  cat("number of x <= mu: ", blow, "\n")
  cat("p.value: ", 2*min(pbinom(bhigh,n,1/2),pbinom(blow,n,1/2)), "\n")

  z <- sort(x)
  alpha <- 1 - conf.level
  k <- qbinom(alpha / 2, n, 1 / 2)
  if (k == 0) k <- k + 1
  cat("sample median: ", median(z), "\n")
  cat("achieved confidence level: ", 1-2*pbinom(k-1,n,1/2), "\n")
  cat("confidence interval for median: ", c(z[k], z[n+1-k]), "\n")
}

> sign.test(na.omit(diff.log10.ifng))
number of x >= mu: 73
number of x <= mu: 209
p.value: 5.107705e-16
sample median: -0.3644727
achieved confidence level: 0.9515989
confidence interval for median: -0.4792927 -0.2561085
```

5. Two-sample tests.

(a) Two-sample *t*-test.

- Parameter of interest: difference in population means ($\mu_1 - \mu_2$).
- Hypotheses:
 - $H_0 : \mu_1 = \mu_2$.
 - $H_1 : \mu_1 \neq \mu_2$.
- Test statistic: Difference in means divided by its standard error.
- Assumptions:
 - Independence of two samples.
 - Normality or large samples (CLT).
 - Common variance (?).

• Sample R commands:

```
# two sample t-test without assuming common variance
> t.test(diff.log10.ifng.boys,diff.log10.ifng.girls)
```

```
Welch Two Sample t-test
```

```
data: diff.log10.ifng.boys and diff.log10.ifng.girls
t = -0.2301, df = 257.594, p-value = 0.8182
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.1636799 0.1294284
sample estimates:
mean of x mean of y
-0.3569724 -0.3398467
```

```
# two sample t-test assuming common variance
> t.test(diff.log10.ifng.boys,diff.log10.ifng.girls,var.equal=T)
```

```
Two Sample t-test
```

```
data: diff.log10.ifng.boys and diff.log10.ifng.girls
t = -0.2308, df = 278, p-value = 0.8177
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.1632245 0.1289731
sample estimates:
mean of x mean of y
-0.3569724 -0.3398467
```

(b) Wilcoxon rank sum test.

- Parameter of interest: location shift (μ).
- Hypotheses:
 - $H_0 : \mu = 0$.
 - $H_1 : \mu \neq 0$.
- Test statistic: Sum of ranks (based on both sample) in the first sample.
- Assumptions:
 - Independence of two samples.
 - Location shift model: $G(x) = F(x - \mu)$ for some μ (CI only).

• Sample R commands:

```
# Wilcoxon rank sum test
> wilcox.test(diff.log10.ifng.boys, diff.log10.ifng.girls, correct=F)
```

```
Wilcoxon rank sum test
```

```
data: diff.log10.ifng.boys and diff.log10.ifng.girls
W = 9803.5, p-value = 0.8054
alternative hypothesis: true mu is not equal to 0
```

```
# Wilcoxon rank sum test with CI for location shift
> wilcox.test(diff.log10.ifng.boys, diff.log10.ifng.girls,
              correct=F, conf.int=T)
```

```
Wilcoxon rank sum test
```

```
data: diff.log10.ifng.boys and diff.log10.ifng.girls
W = 9803.5, p-value = 0.8054
alternative hypothesis: true mu is not equal to 0
95 percent confidence interval:
-0.1253609 0.1662051
sample estimates:
difference in location
0.02014310
```