BIOS 662 Fall 2018

Statistical Inference: Populations and Samples

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Random Variables

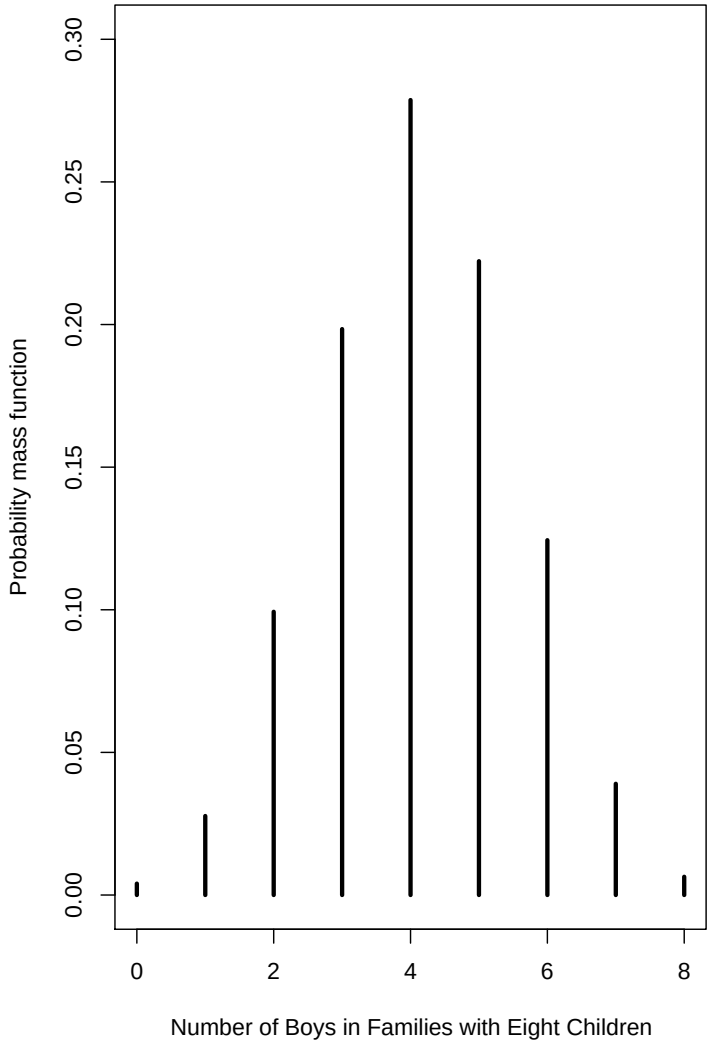
- *Random sample*: result of independently selecting elements at random from a population
- Definition 4.8. A *random variable* is a variable associated with a random sample

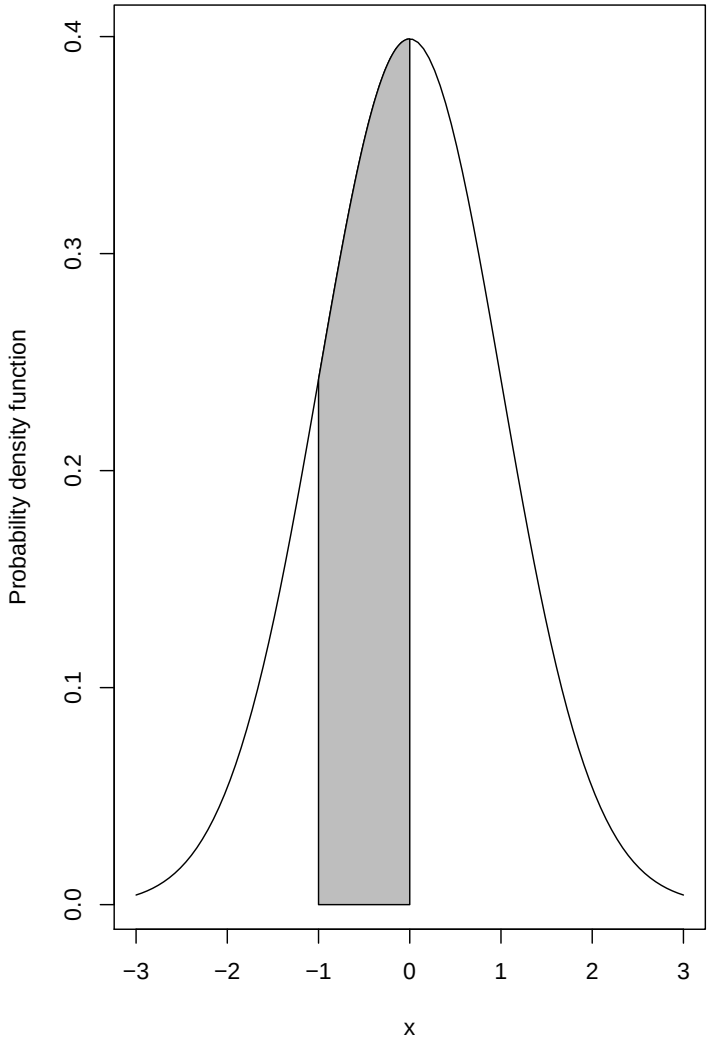
  P.V. Rao (1998, p 786): A *random variable* is a variable whose value is determined by the observed characteristics of an item randomly selected from a population

Probability Functions

- Definition 4.9. The *probability mass function* (pmf) is a function that for each possible value of a discrete random variable takes on the probability of that value occurring
- Definition 4.10. The *probability density function* (pdf) is a curve that specifies, by means of the area under the curve over an interval, the probability that a continuous random variable falls within the interval

Probability Functions





Cumulative Distribution Function

- Definition 4.9. The *cumulative distribution function* for a random variable X is

$$F(x) = \Pr[X \leq x]$$

- If X is discrete,

$$F(x) = \sum_{y \leq x} p_X(y)$$

where p_X is the pmf of X

- If X is continuous,

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

where f is the pdf of X

Population Quantile

- Intuitive definition:

  The p^{th} quantile of X, say ζ_p, should be such that

$$F(\zeta_p) = \Pr[X \leq \zeta_p] = p$$

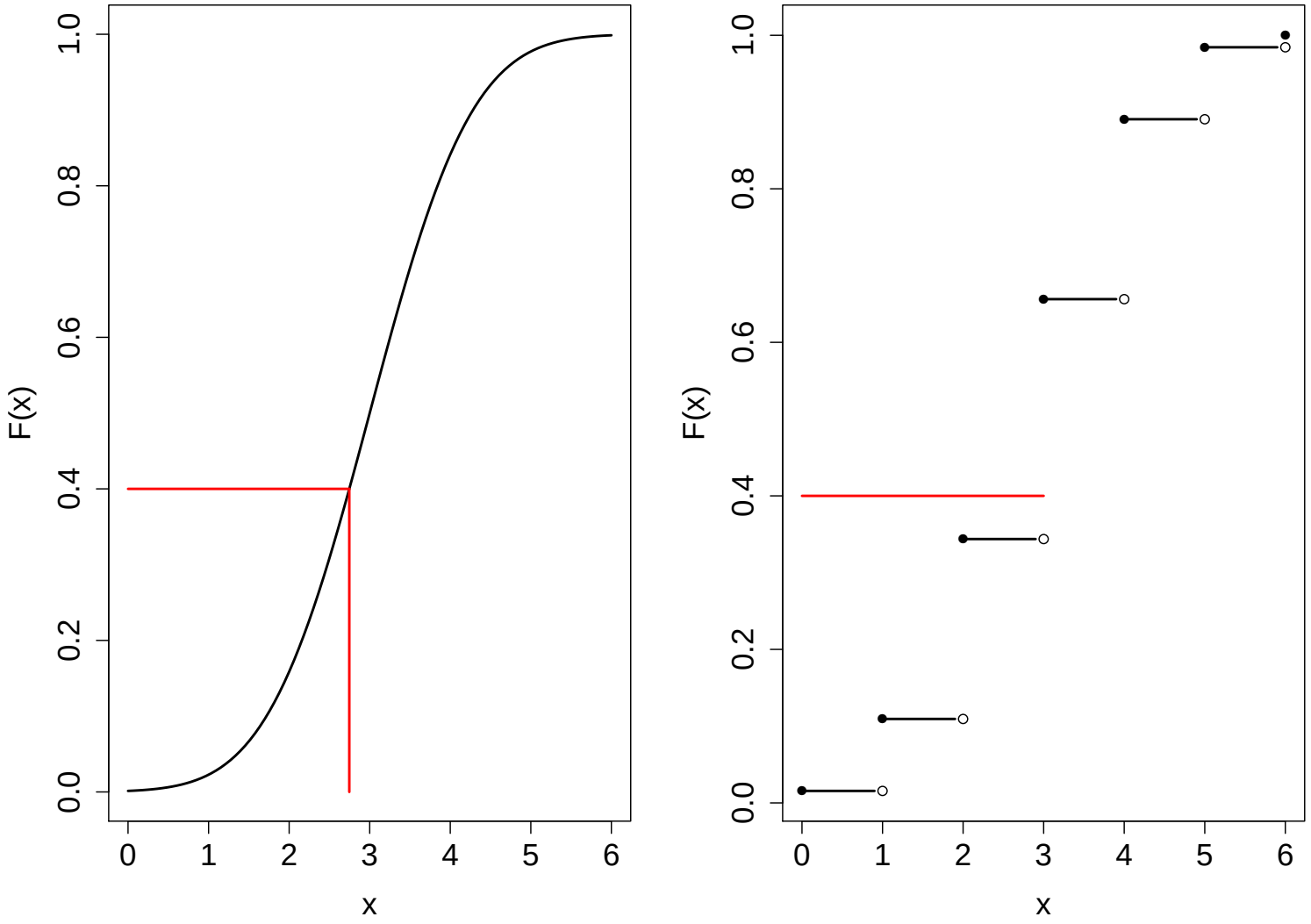
- Formally:

$$\zeta_p = \inf\{x : F(x) \geq p\}$$

- If F is continuous

$$F(\zeta_p) = p$$

Quantiles: Example



Mean and Variance

- Mean or expected value of X
  - If X is discrete,

$$\mu = E(X) = \sum_x x p_X(x)$$

  - If X is continuous,

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- Variance

$$\sigma^2 = \mathrm{Var}(X) = E\{(X-\mu)^2\}$$

- $E(g(X)) = \sum_x g(x) p_X(x)$ (discrete)
  $= \int_{-\infty}^{\infty} g(x) f(x) dx$ (continuous)

Skewness and Kurtosis

- Skewness

$$\alpha_3 = \frac{E\{(X-\mu)^3\}}{\sigma^3}$$

- Kurtosis

$$\alpha_4 = \frac{E\{(X-\mu)^4\}}{\sigma^4}$$

Parameters and Statistics

- Definition: A *parameter* is a numerical characteristic of a population
- Definition: A *statistic* is a numerical characteristic of a sample
- Notation: Greek letters typically denote parameters; Latin / English letters denote statistics
- Example:

  μ population mean; σ^2 population variance

  $\bar{Y}$ sample mean; s^2 sample variance

Parameters and Statistics

- Parameters are fixed constants
- Statistics are random variables
- Statistics have probability distributions
- We will use statistics and probability theory to draw conclusions (inference) about parameters

Sampling Distributions

- Definition 4.15. The probability function of a statistic is called the *sampling distribution of the statistic*
- For example, when sampling from a population, the sample mean $\bar{Y}$ is a random variable because its value depends on chance, namely, on which sample is obtained
- The probability distribution of the random variable $\bar{Y}$ is called the *sampling distribution of the mean*

Sampling Distributions

- Result 4.1. If a random variable Y has a population mean μ and a population variance σ^2, the sampling distribution of the mean ($\bar{Y}$) has mean μ and variance σ^2/n
- Definition 4.16. The standard deviation of the sampling distribution is called the *standard error*
- For example, the standard error of $\bar{Y}$ is $\sigma/\sqrt{n}$

Normal or Gaussian Distribution

- PDF:

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

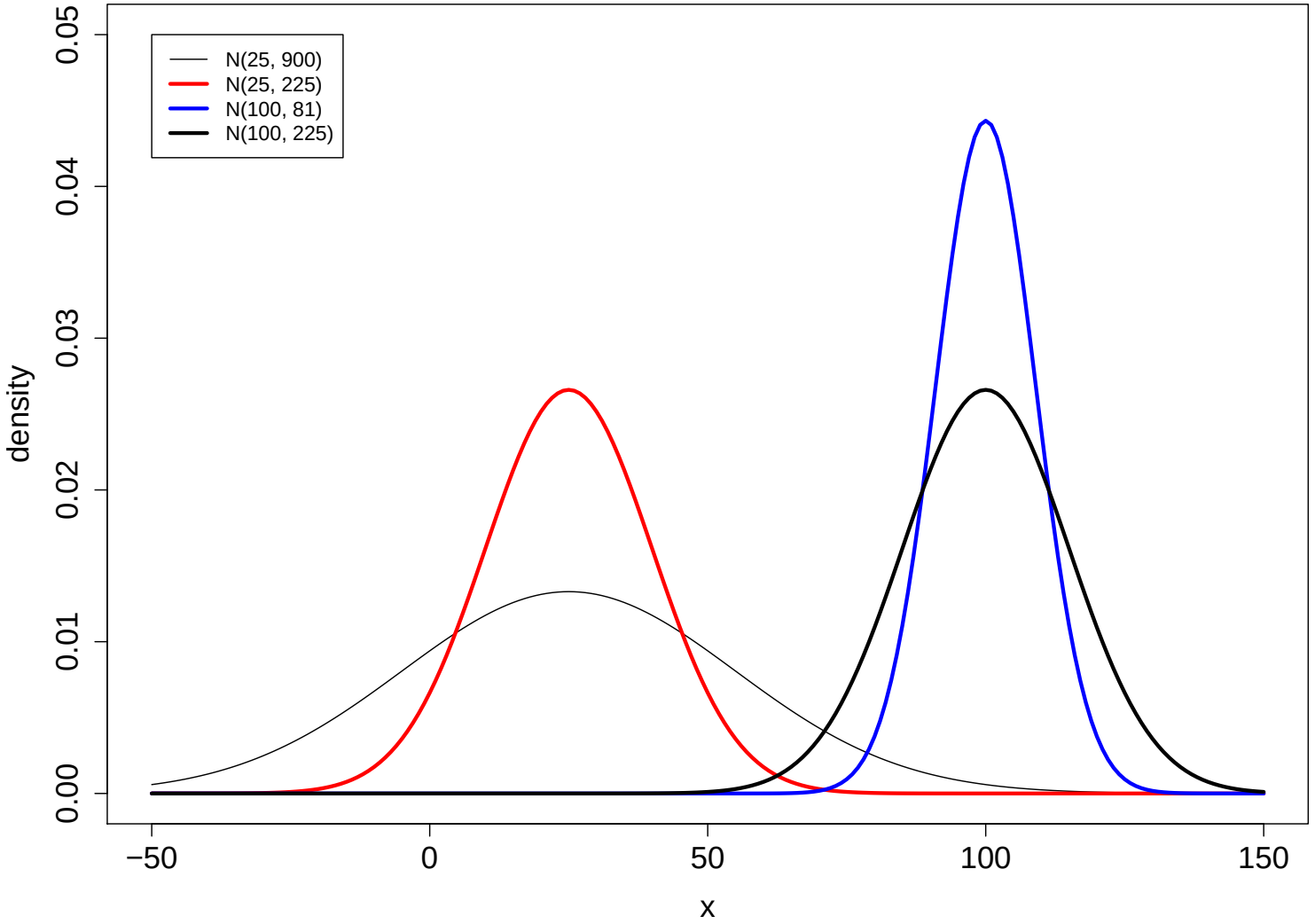
- CDF:

$$F(x;\mu,\sigma) = \int_{-\infty}^{x} f(y;\mu,\sigma)dy$$

- μ mean, σ^2 variance

- $X \sim N(\mu,\sigma^2)$ [beware $X \sim N(\mu,\sigma)$]

Normal Distribution



Standard Normal Distribution

- $Z \sim N(0,1)$
- PDF:

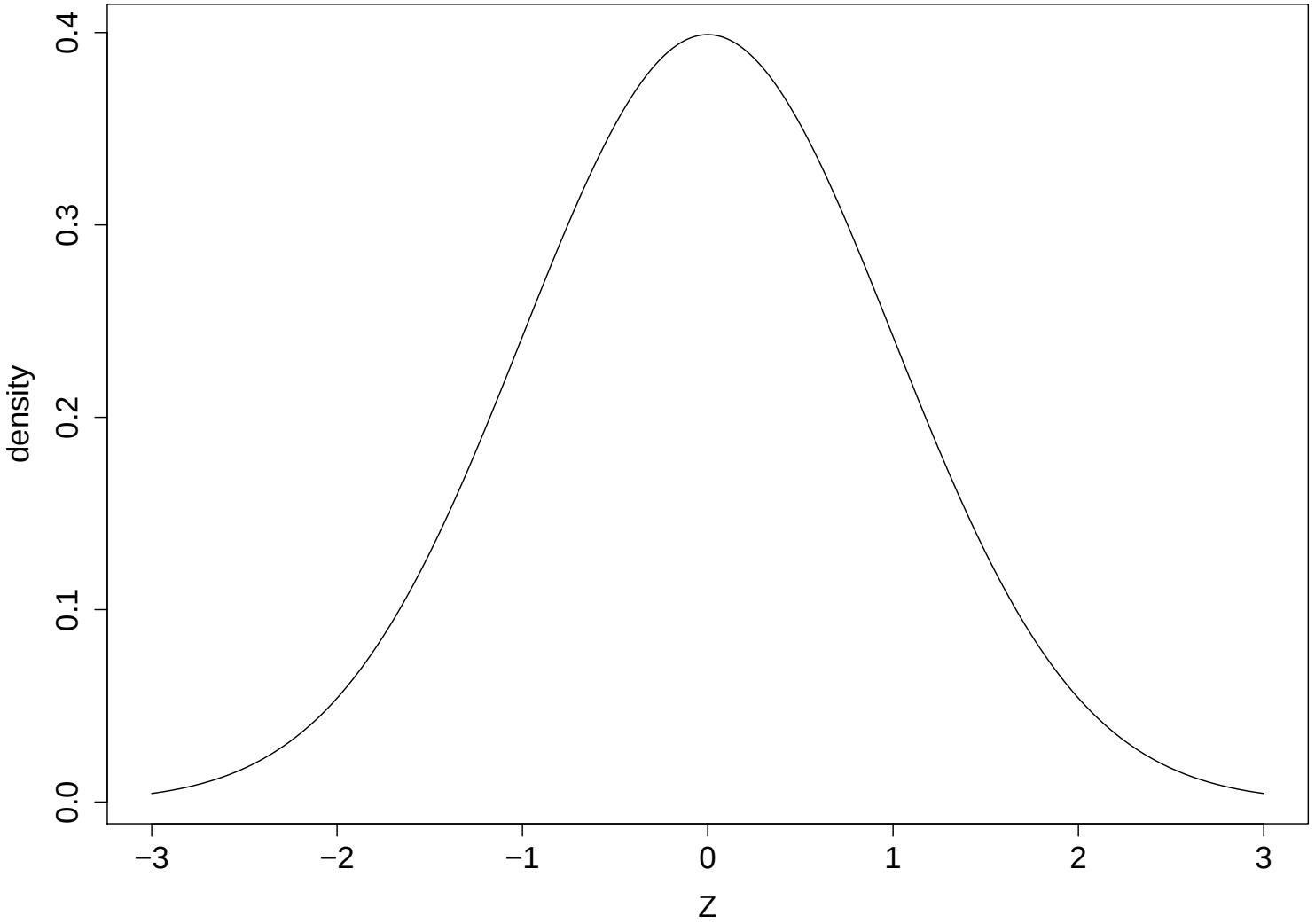
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}z^2\}$$

- CDF:

$$\Phi(z) = \int_{-\infty}^{z} \phi(y)dy$$

- $N(0,1)$ is the *standard normal distribution*

Standard Normal Distribution



Properties of the Standard Normal Distribution

- A random variable with pdf f is *symmetric* about μ if

$$f(\mu + x) = f(\mu - x) \text{ for all } x$$

- $Z \sim N(0,1)$ is symmetric about 0

$$\phi(z) = \phi(-z) \text{ for all } -\infty < z < \infty$$

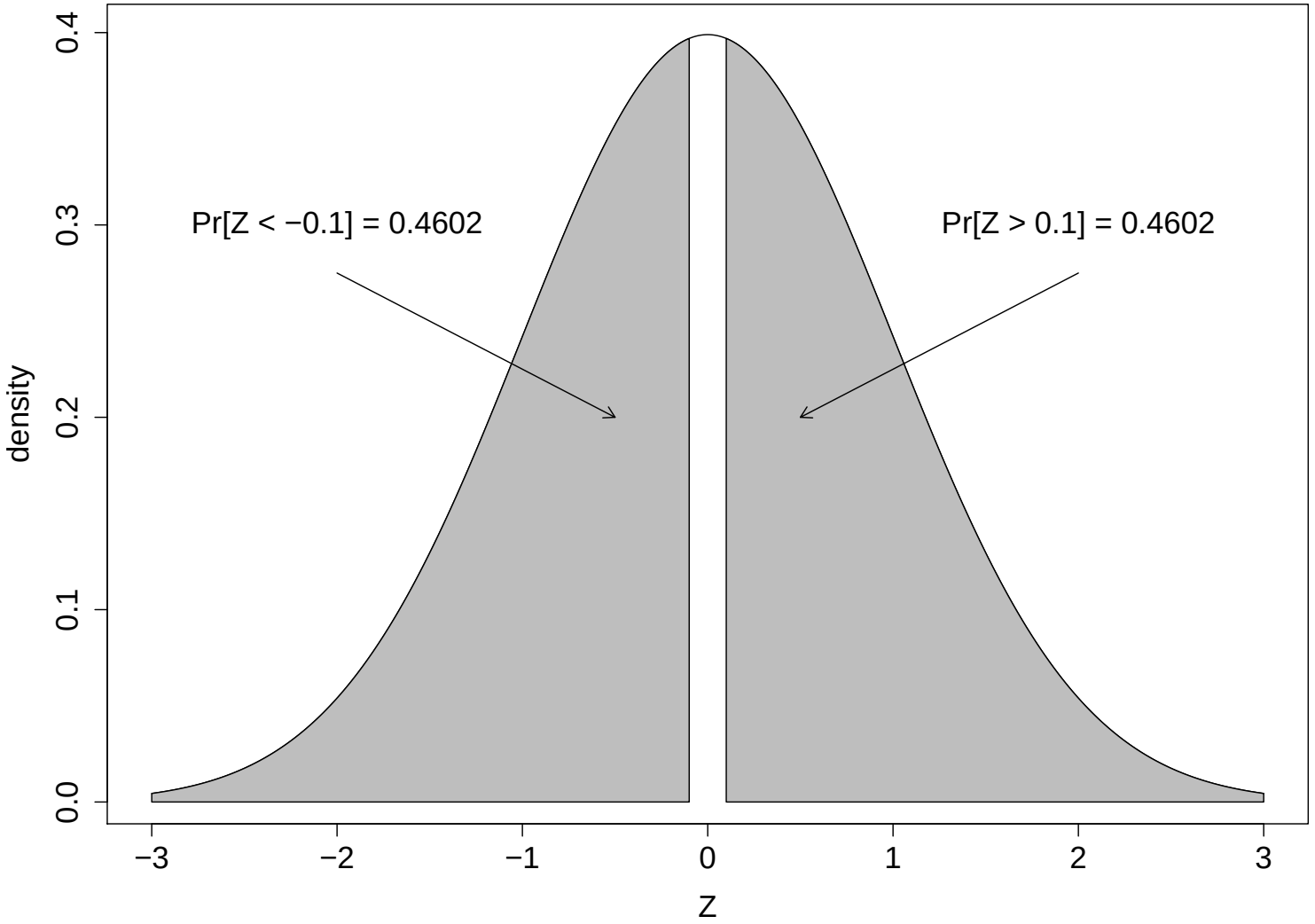
- Thus

$$\Pr[Z \leq -z] = \Pr[Z \geq z]$$

  i.e.

$$\Phi(-z) = 1 - \Phi(z)$$

Standard Normal Distribution



Standard Normal Distribution

- R

```
> pnorm(-0.1,0,1)
[1] 0.4601722
> 1-pnorm(0.1,0,1)
[1] 0.4601722
>  qnorm(0.4601722,0,1)
[1] -0.0999999
```

Standard Normal Distribution

- SAS

```
data normal;
  x=probnorm(-0.1);
  y=cdf('NORMAL',-0.1,0,1);
  z=quantile('NORMAL',0.4601722);

proc print data=normal;
```

Obs	x	y	z
1	0.46017	0.46017	-0.100000

Properties of a Random Variable

- Let X be a random variable
- Suppose

$$Y = aX + b$$

  where a and b are constants

- Then

$$E(Y) = aE(X) + b$$
$$\text{Var}(Y) = a^2 \text{Var}(X)$$

- If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then

$$Y \sim N(a\mu + b, (a\sigma)^2) = N(a\mu + b, a^2\sigma^2)$$

Conversion to Standard Normal

- Suppose $Y \sim N(\mu, \sigma^2)$
- Let

$$Z = \frac{Y - \mu}{\sigma}$$

- Then

$$Z \sim N(0, 1)$$

- In words: any normally distributed random variable can be standardized by subtracting its mean and dividing by its standard deviation

Computation of Probabilities

- Suppose $Y \sim N(\mu, \sigma^2)$
- Let

$$Z = \frac{Y - \mu}{\sigma}$$

- Then

$$\Pr[a < Y < b] = \Pr[\tfrac{a-\mu}{\sigma} < Z < \tfrac{b-\mu}{\sigma}]$$

$$= \Phi(\tfrac{b-\mu}{\sigma}) - \Phi(\tfrac{a-\mu}{\sigma})$$

Table 1 (text, p 818): Standard Normal Distribution

Let Z be a normal random variable with mean zero and variance one. For selected values of z, three values are tabled: (1) the two-sided p-value, or $\Pr[|Z| \geq z]$; (2) the one-sided p-value, or $\Pr[Z \geq z]$; and (3) the cumulative distribution function at z, or $\Pr[Z \leq z]$.

z	Two-sided	One-sided	Cum-dist.
0.00	1.0000	.5000	.5000
0.05	.9601	.4801	.5199
0.10	.9203	.4602	.5398
0.15	.8808	.4404	.5596
0.20	.8415	.4207	.5793
0.25	.8026	.4013	.5987
0.30	.7642	.3821	.6179
0.35	.7263	.3632	.6368
0.40	.6892	.3446	.6554
0.45	.6527	.3264	.6736
⋮			
1.00	.3173	.1587	.8413
1.33	.1835	.0918	.9082
1.64	.1010	.0505	.9495
1.96	.0500	.0250	.9750
2.00	.0455	.0288	.9772
2.58	.0099	.0049	.9951

Example

- Intraocular pressure (IP) is used to diagnose glaucoma
- Assume IP is normally distributed with
  mean $\mu = 16$ mmHg
  and variance $\sigma^2 = 9$ mmHg^2
- If pressure greater than 20 mmHg is considered abnormal, what proportion of the population is abnormal?

$$\begin{aligned}\Pr[X > 20] &= \Pr[\tfrac{X-16}{3} > \tfrac{20-16}{3}] \\ &= \Pr[Z > 1.33] = 1 - \Phi(1.33) \\ &= 1 - 0.9082 = 0.0918\end{aligned}$$

Example (continued)

- What proportion of the population has IP between 4 and 18?

$$\begin{aligned}
\Pr[4 < X < 18] &= \Pr[\tfrac{4-16}{3} < \tfrac{X-16}{3} < \tfrac{18-16}{3}] \\
&= \Pr[-4 < Z < 2/3] \\
&= \Phi(2/3) - \Phi(-4) \\
&= \Phi(2/3) - 1 + \Phi(4) \\
&= 0.7475 - 1 + 0.99997 = 0.7475
\end{aligned}$$

Assessing Normality

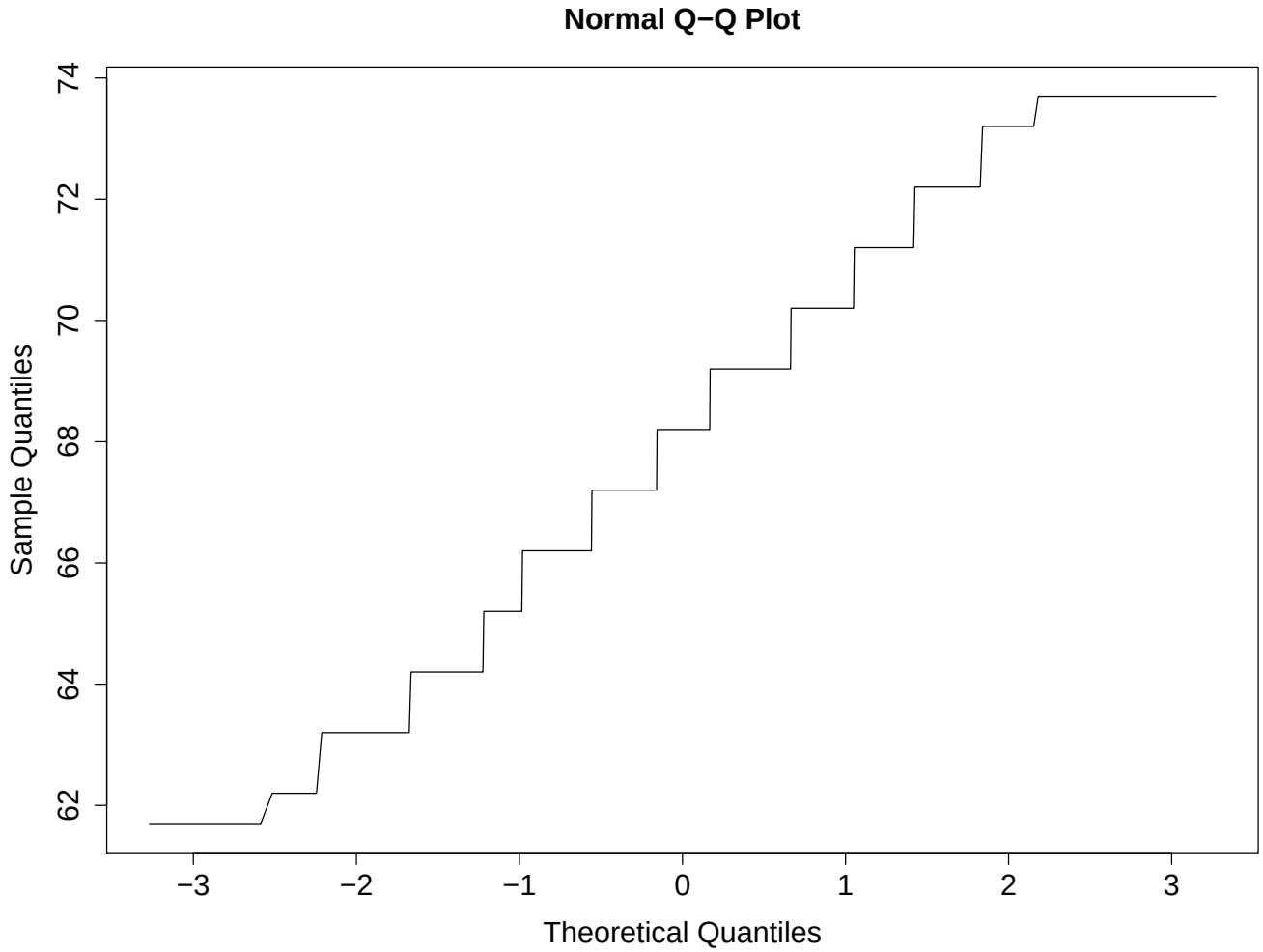
- How do we assess whether the normal distribution model is a reasonable fit for a particular set of data?
- One graphical approach: quantile-quantile (QQ) plot
- Plot quantiles of the observed data distribution versus the quantiles of the normal distribution
- Straight line indicates normality assumption reasonable

QQ Plot Example

- Table 4.3 from text (p. 81)

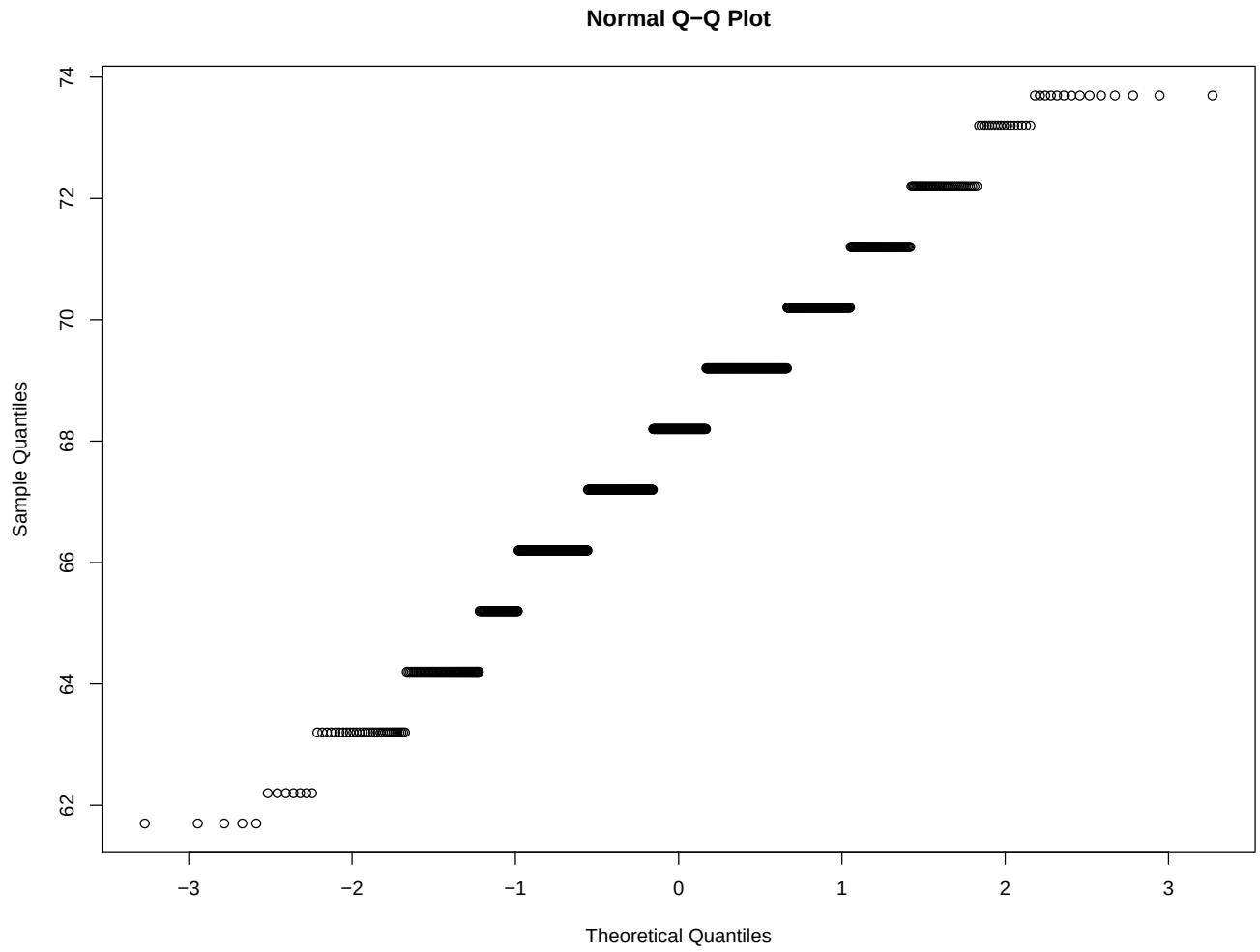
Endpoint	Frequency	Cumulative Percentage
61.7	5	0.5
62.2	7	1.3
63.2	32	4.7
64.2	59	11.1
65.2	48	16.3
66.2	117	28.9
67.2	138	43.8
68.2	120	56.7
69.2	167	74.7
70.2	99	85.3
71.2	64	92.2
72.2	41	96.7
73.2	17	98.5
73.7	14	100.0

R QQ Plot Example: Table 4.3 from Text



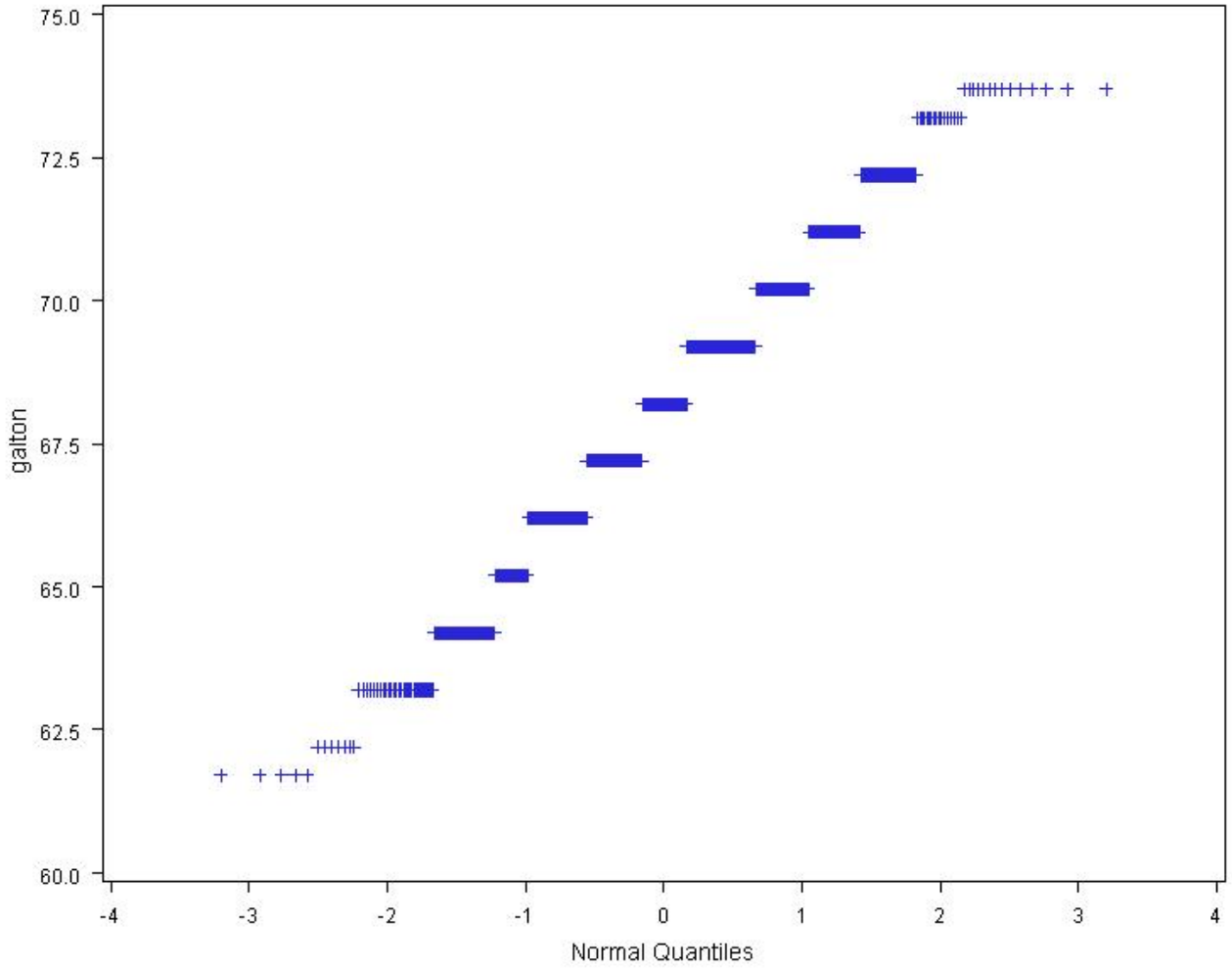

```
> qqnorm(galton,type="l")    # type="l" draws lines
```

R QQ Plot Example: Table 4.3 from Text



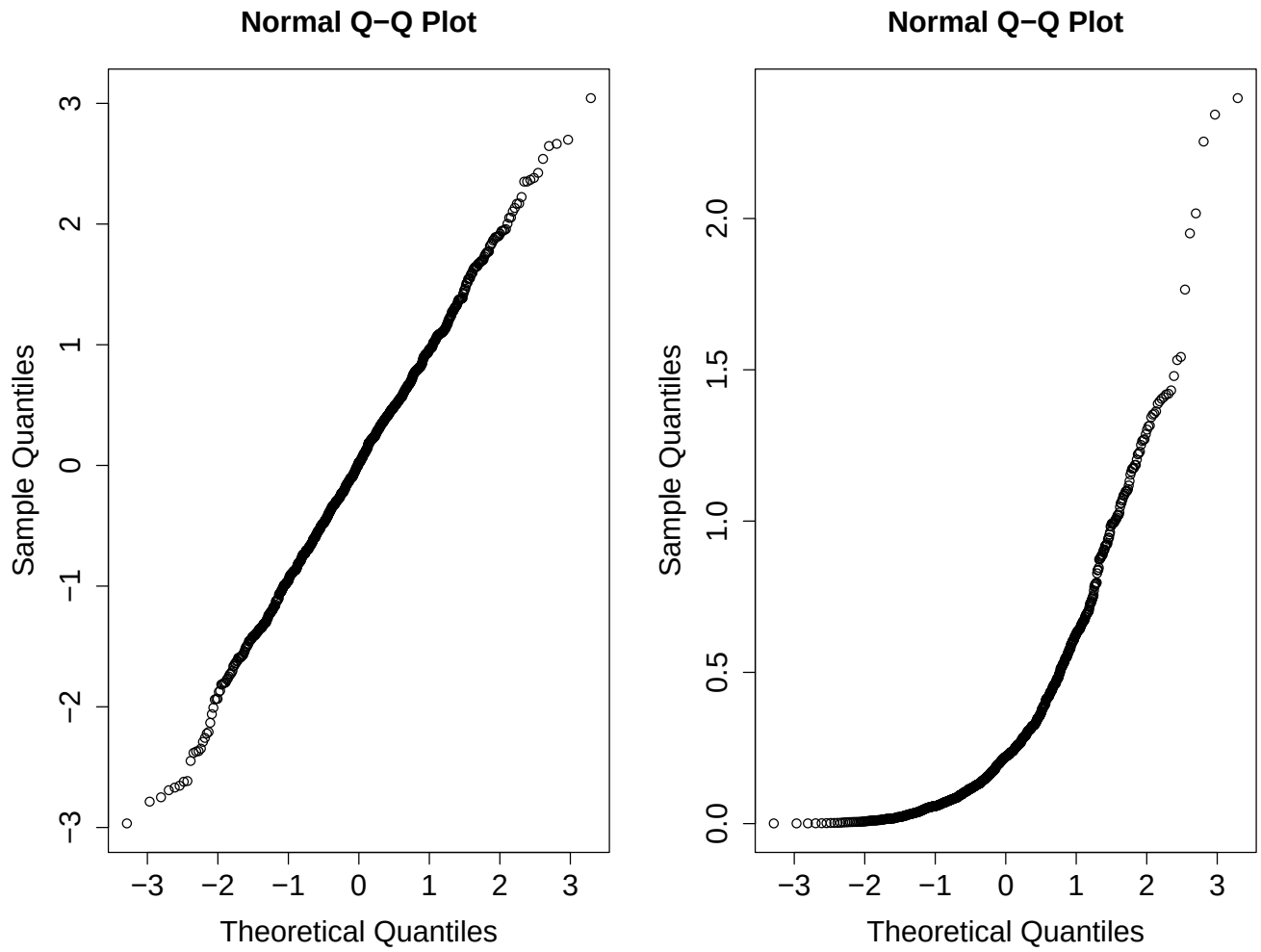

```
> qqnorm(galton)
```

SAS QQ Plot Example: Table 4.3 from Text



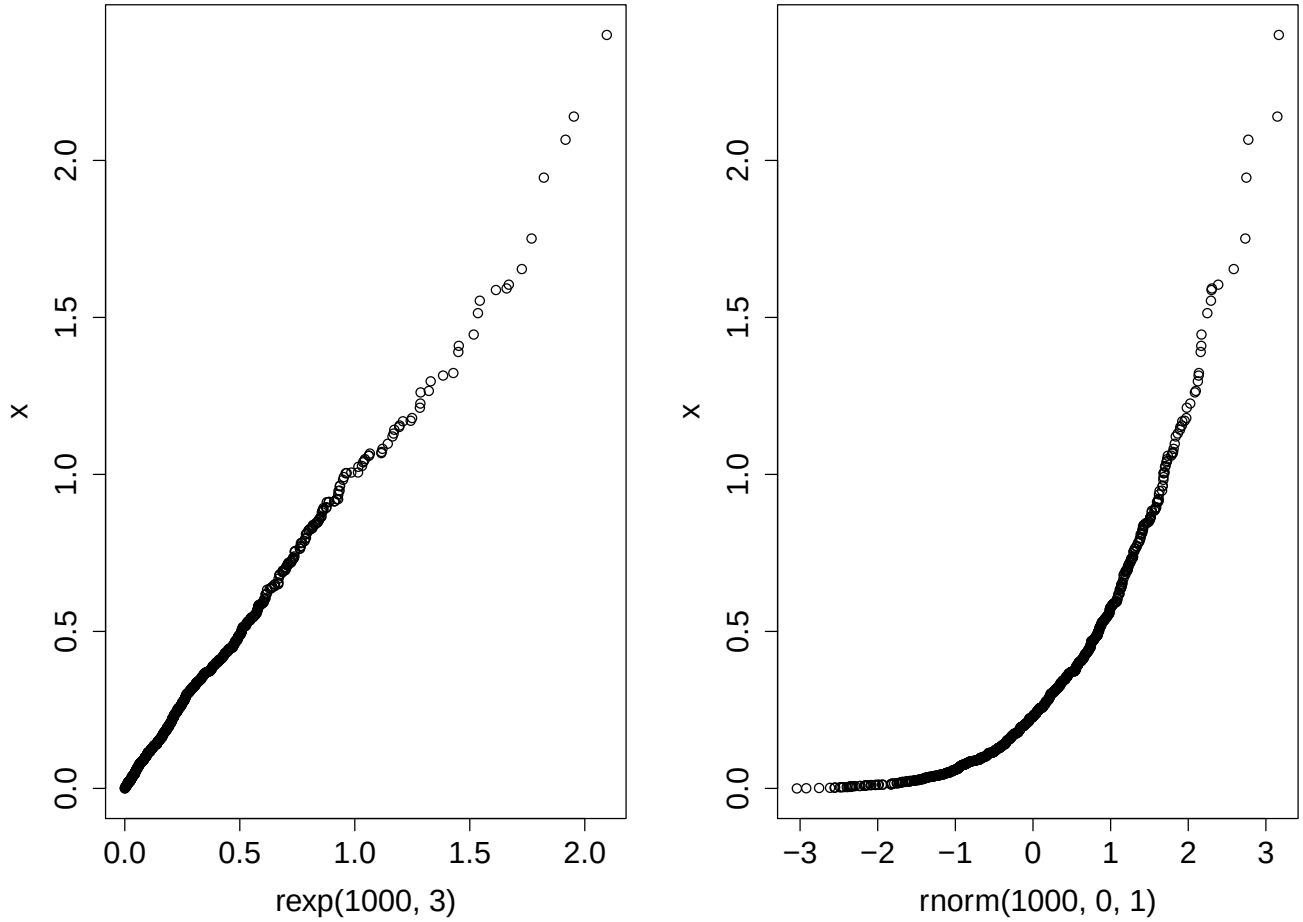

```
proc univariate;
   qqplot galton;
```

R QQ Plots



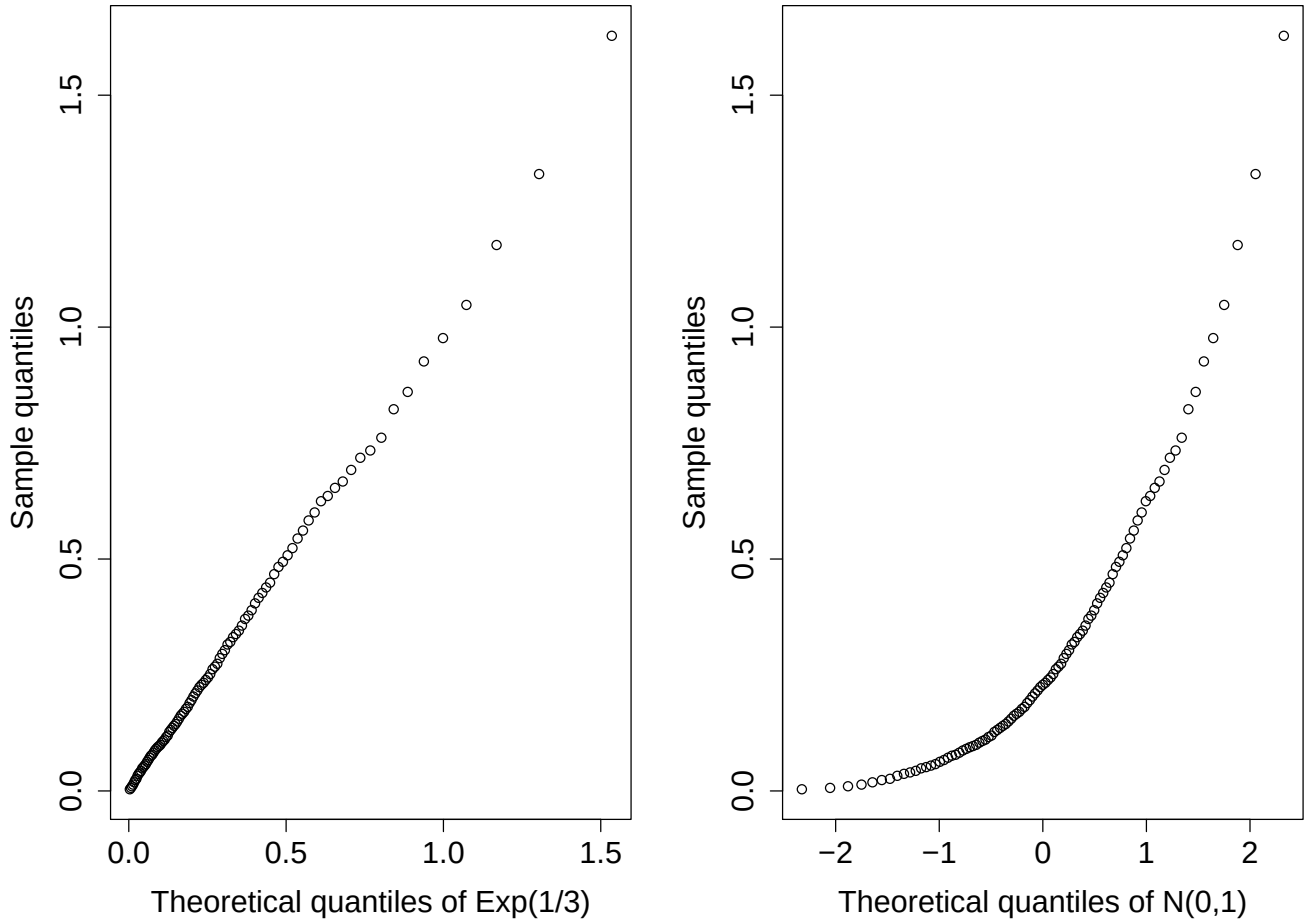

```
> par(mfcol=c(1,2)); qqnorm(rnorm(1000,0,1)); qqnorm(rexp(1000,3))
```

R QQ Plots




```
> x <- rexp(1000,3)
> par(mfcol=c(1,2)); qqplot(rexp(1000,3),x); qqplot(rnorm(1000,0,1),x)
```

R QQ Plots




```
> x <- rexp(1000,3); probs <- seq(0.01,0.99,length=99); par(mfcol=c(1,2))
> qx <- quantile(x,probs); tqexp <- qexp(probs,3); tqnorm <- qnorm(probs,0,1)
> plot(tqexp,qx,xlab="Theoretical quantiles of Exp(1/3)", ylab="Sample quantiles")
> plot(tqnorm,qx,xlab="Theoretical quantiles of N(0,1)",ylab="Sample quantiles")
```

Some Approximations for the Normal

- The interval $\bar{x} \pm s$ will contain approx 68% of the observations
- The interval $\bar{x} \pm 2s$ will contain approx 95% of the observations
- Assuming $Y \sim N(\mu, \sigma^2)$

$$\Pr[\mu - \sigma < Y < \mu + \sigma] = \Pr[-1 < Z < 1] = 0.6827$$

$$\Pr[\mu - 2\sigma < Y < \mu + 2\sigma] = \Pr[-2 < Z < 2] = 0.9545$$

Some Approximations

- Do these approximations hold for non-normal data?

  Not in general.

- Consider $X \sim \text{Exp}(1/\lambda)$ such that $E(X) = \lambda$ and $\text{Var}(X) = \lambda^2$.

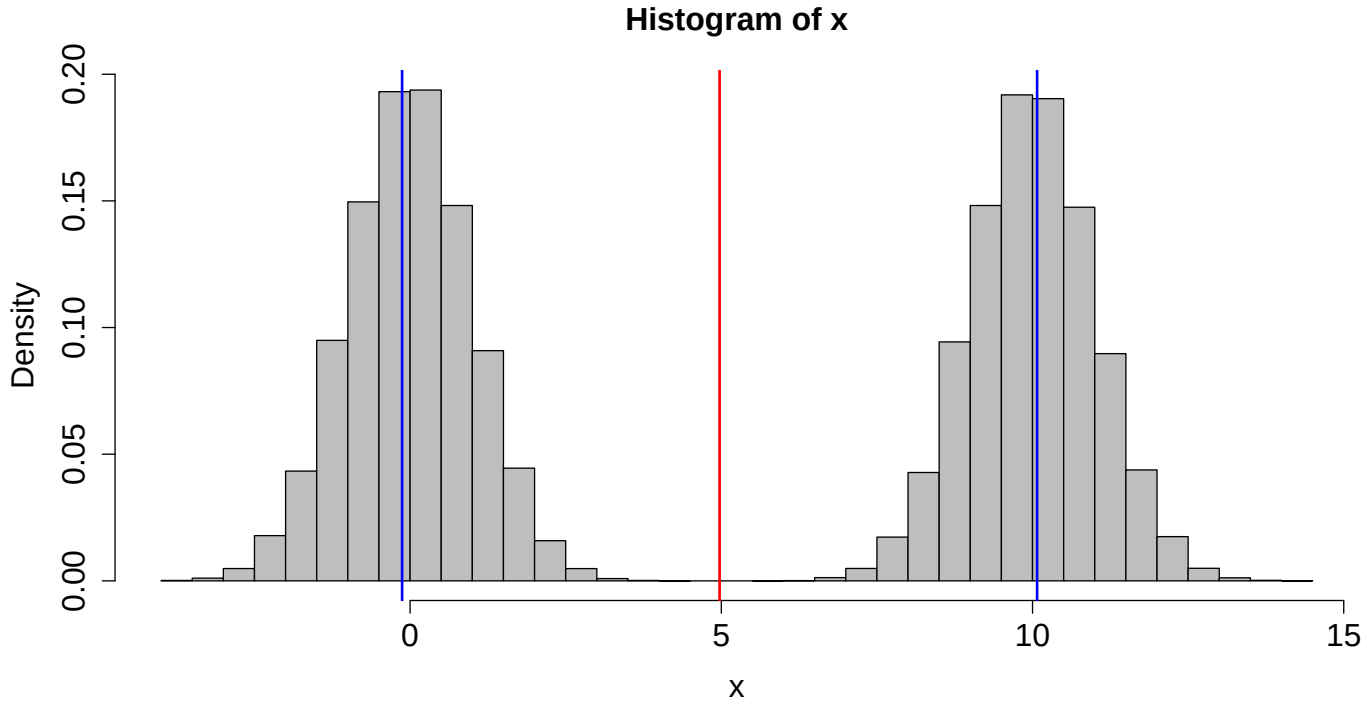
  For $\lambda = 1/3$,

$$\Pr[0 \leq X \leq 2/3] = 0.86$$

Some Approximations

- Consider $X = WY + (1-W)Z$
  where $W \sim \text{Bernoulli}(1/2)$, $Y \sim N(10,1)$,
  and $Z \sim N(0,1)$.

  Can show $\Pr[\mu_X - \sigma_X \le X \le \mu_X + \sigma_X] \approx 0.54$



Some Approximations

- The following holds for any random variable Y with mean μ and variance σ^2

$$\Pr\left[\left|\frac{Y-\mu}{\sigma}\right| < K\right] = \Pr\left[\mu - K\sigma < Y < \mu + K\sigma\right] \geq 1 - \frac{1}{K^2}$$

  $\forall K \geq 1$. This is *Chebyshev's inequality* (note typos in text on page 100)

- For example, if $K = 2$,

$$\Pr\left[\mu - 2\sigma < Y < \mu + 2\sigma\right] \geq 0.75$$

  i.e. we would expect at least 75% of observations to be within two standard deviations of the mean *for any underlying distribution*

Central Limit Theorem (CLT)

- Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed (iid) random variables with

$$E(Y_i) = \mu \quad \text{(finite)}$$

and

$$\text{Var}(Y_i) = \sigma^2 > 0$$

- Define

$$Z_n = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

- Then the distribution function of Z_n converges to the standard normal distribution function as $n \to \infty$.

Central Limit Theorem (CLT)

- In words - see Result 4.3 on page 84 of the textbook
- If a random variable Y has population mean μ and population variance σ^2, then the sample mean $\bar{Y}$, based on n observations, is approximately normally distributed with mean μ and variance σ^2/n for sufficiently large n

Notes on the CLT

- The CLT applies to any distribution of the Ys
- The approximation improves as n gets large
- Check out the Rice Virtual Lab in Statistics

http://onlinestatbook.com/rvls.html

Result 4.2

- If Y is *normally distributed* with mean μ and variance σ^2, then $\bar{Y}$, based on a random sample of n observations, is normally distributed with mean μ and variance σ^2/n.
- This is true regardless of sample size.